UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE CIENCIAS FÍSICAS
Departamento de Física Atómica, Nuclear y Molecular

LA AUTORREGULACIÓN TRANSCRIPCIONAL EN LA OPTIMIZACIÓN Y ENSAMBLAJE DE LOS MOTIVOS DE RED BACTERIANOS.

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Francisco Miguel Camas Gallego

Bajo la dirección de los doctores

Juan Fernando Poyatos Adeva
Jesús Blázquez Gómez

Madrid, 2010

- **ISBN: 978-84-693-2389-2**

Tesis doctoral

La autorregulación transcripcional en la optimización y ensamblaje de los motivos de red bacterianos

Francisco Miguel Camas Gallego

Programa de doctorado:
Física de Sistemas Complejos
Departamento de Física Atómica, Nuclear y Molecular
Universidad Complutense de Madrid

Codirectores:
Juan Fernando Poyatos Adeva
Jesús Blázquez Gómez

Madrid, Mayo 2009

A mi mujer, Teresa

Agradecimientos

La lista de personas a las que tengo que agradecer que, en mi modesta opinión, esta tesis haya llegado a buen puerto ha de empezar inexcusablemente por mis directores en el Centro Nacional de Biotecnología, Juan F. Poyatos y Jesús Blázquez, los cuales, yendo contracorriente de los usos ordinarios y con el propósito de hacer de mí un científico de la máxima autonomía posible, me dieron toda la libertad para perderme y reencontrarme por los mil caminos de la investigación, asumiendo los riesgos que ello conlleva; pues aunque "la libertad, Sancho, es uno de los más preciosos dones que a los hombres dieron los cielos", la osadía de la inexperiencia también puede dar lugar a entuertos que no se puedan desfacer. He de agradecer asimismo a mi tutor de la Universidad Complutense de Madrid, Juan M. R. Parrondo, su labor de mediación con esta institución y su espléndido curso de Fluctuaciones en Sistemas Complejos.

Una estancia en el Centro de Astrobiología en el tramo final de la licenciatura marcó definitivamente el rumbo de mi carrera científica y propició mi transición de la Física al mundo de los bichos. En este sentido, el papel jugado por José María Gómez Gómez, el indefinible Chema, ha sido fundamental. Enamorado de la ciencia hasta la locura, científico romántico y humanista, constructor de zigurats astronómicos, su descubrimiento no reconocido del gen *hnr* es una muestra de las injusticias que puede llegar a amparar la trastienda del mundo de la investigación. También del CAB es heredera mi amistad con Antonio Salgado, que es el contrapeso apolíneo del anterior, y que fue la persona con la que empezó mi tránsito del largo trecho que va de estudiante de ciencias a científico. La experiencia de haber conocido a Chema y a Antonio basta para que haya merecido la pena aparcar las partituras durante estos últimos años.

Gracias a su capacidad para mantener la ilusión científica por encima de los avatares del momento, mis compañeros del Centro Nacional de Biotecnología, Alejandro Couce, Alexandro R. Rojas y Javier Ramos, han sido fundamentales a la hora de sobrellevar la dureza de los días experimentales que terminan en blanco, que no son pocos. El tramo central de mi doctorado se desarrolló en el Centro Nacional de Investigaciones Oncológicas, donde tuve la fortuna de coincidir con Ramón Díaz-Uriarte y Oscar M. Rueda, de cuyos amplios conocimientos estadísticos esta tesis es deudora. Sin salirnos del CNIO, también he de agradecer a Andrés

Cañada y Andreu Alibés sus buenos consejos informáticos; a Edward Morrissey, su generosidad sin matices a la hora de compartir conocimientos y su fiel amistad; y a Pilar Redondo, lo mucho que facilitó mi trabajo experimental en este centro.

El tiempo que me ahorró con los endiablados ordenadores o el empeño que puso en trastocar el sedentarismo al que soy tan proclive bastarían para hacer constar la gratitud que le debo a Sandra Moreno. Sin embargo, a mi única compañera en el laboratorio de Lógica de los Sistemas Genómicos durante buena parte de la tesis le agradezco, ante todo, precisamente eso: el estar ahí con su vitalidad inagotable, que no es poca hacienda cuando uno ha de enfrentarse con el día a día de la investigación.

En estos agradecimientos no puede faltar el profesor Eric J. Alm, que me acogió en su laboratorio en el Instituto de Tecnología de Massachusetts y cuyo dominio de las bases de datos genómicas puso a mi entera disposición en aras de abordar el código del que nos ocupamos en el tramo final de este documento. De hecho, al margen de las cuestiones técnicas, mi proyecto de resolución de tal código encontró en él el impulso anímico decisivo. A Luis A. Padrón he de agradecerle que me sacara las castañas del fuego burocrático en más de una ocasión (y en más de un continente) y que haya sido en buena medida el responsable de que mi estancia primeriza en Boston fuese más que grata.

Siento la más profunda gratitud hacia mis padres, Miguel y Carmen, de orígenes muy humildes, porque siempre tuvieron claro que la educación era el camino, y así, nunca escatimaron esfuerzos en este sentido, incluso en los tiempos de más estrechez económica, que fueron casi todos. El ánimo emprendedor y el ingenio de mi padre, y la fortaleza y capacidad de sacrificio de mi madre siempre serán referentes que no podré alcanzar. En mis progenitores, en mis hermanos Susana y Fernando, en el resto de mi familia y en mi familia política siempre he encontrado el cariño sobre el que se sustenta todo lo demás.

***

Sin ella, ni la licenciatura en Físicas ni esta tesis hubiesen sido posibles; así que, por encima de todo, quiero agradecer a mi mujer, Teresa, su amor y su entrega, que se han traducido siempre en un apoyo incondicional y sin fisuras a toda empresa que he ido acometiendo. Teresa podría haber tenido una vida mucho menos complicada si, en lugar de embarcarme en una carrera científica, hubiese dado continuidad a la estabilidad alcanzada en el mundo de la enseñanza de la Música. Los sacrificios y privaciones a los que se ha visto sometida debido a mis decisiones suponen una deuda que nunca terminará de estar saldada.

Índice general

Nomenclatura

ADN ácido desoxirribonucleico

AP actividad del promotor

ARN ácido ribonucleico

BS sitio de unión al ADN de un factor de transcripción (por *binding site*)

DO densidad óptica

FFL *feed-forward loop*

GFP proteína de fluorescencia verde (por *green fluorescence protein*)

HR hélice de reconocimiento

HTH hélice-giro-hélice (por *helix-turn-helix*)

IPTG isopropil-β-D-tiogalactopiranósido

mARN ARN mensajero

PF huellas digitales filogenéticas (por *phylogenetic footprinting*)

PSD densidad espectral de potencia (por *power spectral density*)

PWM matriz de peso de las posiciones (por *position weight matrix*)

RMN resonancia magnética nuclear

RTB RegTransBase

SIM módulo de entrada simple (*por single input module*)

TF factor de transcripción (por *transcriptional factor*)

UV ultravioleta

Introducción

Las distintas vertientes de la genómica funcional –secuenciaciones de genomas, transcriptómica, proteómica y metabolómica– suponen el punto culminante de la Biología Molecular en cuanto a la identificación de los componentes celulares y sus concentraciones se refiere [1]. Sin embargo, por no ocuparse de las interacciones entre las moléculas ni de cómo éstas se organizan en la célula, este tipo de inventariado exhaustivo no basta para explicar el funcionamiento de la célula viva, lo cual pone de manifiesto definitivamente las limitaciones del paradigma reduccionista de la Biología Molecular. Este reduccionismo en su definición más radical propugna que para el entendimiento del sistema celular basta con una descripción completa de las propiedades de los componentes moleculares del mismo, sin más añadiduras [2].

Maticemos que lo anterior no significa que, al margen de las *ómicas*, la Biología Molecular no se ocupe de las interacciones; lo que sucede es que la descripción de éstas se detiene en lo meramente cualitativo –adoptando en su plasmación visual la forma de los recurrentes diagramas de "flechas" [3]. Bajo este abordaje cualitativo subyace la presunción más o menos implícita –o más o menos consciente, si se prefiere– de que el comportamiento de los sistemas celulares tiene un carácter aproximadamente lineal, o dicho de otro modo, de que la correlación entre las dinámicas de dos especies moleculares en interacción tiene siempre el mismo signo. Por tanto, por considerar que el comportamiento del sistema se puede inferir a partir de las intuiciones lineales sugeridas por el boceto de las conexiones entre sus componentes, una aproximación de este tipo no necesita despegarse demasiado de la descripción de las propiedades moleculares.

Si la linealidad estuviese realmente en la naturaleza de las interacciones moleculares el todo sería efectivamente poco más que la suma de las partes. Sin embargo, estas interacciones tienen un carácter no lineal, lo cual implica que incluso a partir del establecimiento de relaciones –por ejemplo, de regulación– entre unas pocas especies moleculares puede emerger un comportamiento de gran complejidad que trascienda con mucho el esperable del análisis por separado de dichas especies. De hecho, un mismo mapa de conexiones puede dar lugar a un repertorio de fenotipos dinámicos bien distintos [4]. La identificación de este repertorio está muchas veces lejos del alcance de la intuición cualitativa, con lo que

la dinámica emergente sólo puede ser resuelta en términos matemáticos.

Además, tanto el preciso régimen dinámico en el que, de entre todas las opciones potenciales, se encuentra el sistema, como aquellos otros regímenes que éste tiene fisiológicamente a su alcance vienen determinados por las concentraciones de las distintas especies en interacción (las variables del sistema) y por el valor de los parámetros que rigen dichas interacciones. Por tanto, la resolución de la funcionalidad de un sistema biológico requiere que la aproximación matemática vaya acompañada de la cuantificación de estas magnitudes.

Por todo esto, la Biología está efectuando una transición hacia estudios más cuantitativos bajo la premisa de que la fisiología de la célula es fruto de interacciones de carácter no-lineal entre subconjuntos de componentes moleculares de ésta. La nueva aproximación, que fija su objeto de estudio en niveles de organización que van más allá de los ladrillos moleculares –y que lo hace de una forma eminentemente cuantitativa, combinando los experimentos *in vivo* con la modelización matemática del sistema en cuestión– es lo que se ha dado en llamar Biología de Sistemas, disciplina que tiene visos de desbancar a la Biología Molecular como corriente dominante dentro de las ciencias de la vida.

Las limitaciones del paradigma reduccionista no implican que la Biología de Sistemas haya adoptado una posición holística radical –bajo la cual el sistema celular sólo podría ser entendido como un todo irreducible y en términos de la mera fenomenología fisiológica [2]. La razón de que la nueva ciencia se haya podido situar a medio camino de uno y otro extremo reside en la presumible naturaleza modular de las funciones celulares –un módulo consistiría en un subconjunto delimitado de especies moleculares de cuya interacción emerge una funcionalidad separable hasta cierto grado de la del resto de módulos [5]. Ello permite que la emergencia de las propiedades fisiológicas pueda estudiarse a partir de la reducción del sistema a niveles de organización intermedios.

En realidad, esta descripción del método de trabajo de la Biología de Sistemas es sólo una de las dos aproximaciones que esta disciplina ha adoptado: la que va "de abajo a arriba" (*bottom-up*) infiriendo el comportamiento del sistema a partir de las interacciones de sus componentes. La otra postura, "de arriba a abajo" (*top-down*), parte precisamente del estudio de las correlaciones que se puedan observar en la ingente cantidad de datos generados por las disciplinas ómicas, para intentar deducir la manera en que se organizan las funciones celulares [1].

La primera de estas aproximaciones tiene su origen en la adopción de conceptos y formalismos provenientes de la Teoría de Control, de cuyo corpus teórico (junto a los de la Biofísica y la Termodinámica fuera del equilibrio) es en gran parte deudora la Biología de Sistemas [6]. Por tanto, no resultará extraño que uno de los objetos de estudio de esta disciplina sea la dinámica de los mecanismos de control celulares basados en procesos de realimentación. Quizás el caso más sencillo de éstos sea la autorregulación o control autógeno de la transcripción, que es la regulación de la transcripción de un gen por la propia proteína codificada

en dicho gen. A semejanza de lo que sucede en los sistemas no biológicos con realimentación, se ha vinculado la autorregulación transcripcional, por vía de sus propiedades dinámicas, a la optimización de las funciones celulares [7, 8].

Como hemos mencionado, el conjunto de estas funciones tiende a disponerse en arquitecturas modulares. Sin embargo, la definición de módulo referida más arriba es muy laxa pues en ella caben desde el sistema SOS de respuesta inducible ante daños en el ADN de la bacteria *Escherichia coli* hasta los orgánulos de los eucariotas. Por ello, ante la evidencia de que en la célula existen estos niveles de organización intermedios, surgió en la Biología de Sistemas otra línea de investigación de mucha pujanza que consiste en la búsqueda de posibles principios –más concretos en su definición– que rijan la organización de la arquitectura de las redes moleculares y que llevó a la identificación de patrones de conexión recurrentes –primero en la red de transcripción de *E. coli* [9] y luego en otros organismos– denominados motivos de red [10], entre los cuales se encuentra la propia autorregulación ($X \rightarrow X$) o el *feed-forward loop*, que involucra a tres nodos de la red con las siguientes conexiones: $X \rightarrow Y$, $X \rightarrow Z$, $Y \rightarrow Z$. La relevancia de la señal estadística en la que descansaba la identificación de los motivos llevó a que se los considerara objeto de selección adaptativa [11]. Además, estudios tanto teóricos como experimentales asociaron a estos motivos una serie de propiedades dinámicas que se postularon como la causa por la que, en última instancia, estos motivos –a modo de unidades "ingenieriles" de procesamiento de información– serían seleccionados [10, 12]. No obstante, como veremos, no faltan las posturas críticas –unas frontales y otras de matiz– con este escenario adaptativo [13–16].

En ocasiones es fácil hacer la identificación de algunos de los motivos con los propios módulos funcionales. Así sucede, por ejemplo, en el caso del sistema SOS antes mencionado, cuyo esquema de conexiones transcripcionales dibuja un arquetípico motivo SIM –por *single input module*: $X \rightarrow Y$, $X \rightarrow Y'$, $X \rightarrow Y''$, etc. Además, aunque en muchos casos el número de elementos que los constituyen es demasiado pequeño como para implementar una función celular, los motivos se presentan en forma de agregados en los que dos o más motivos comparten algunos de sus elementos [17,18], como ocurre en el caso de los FFLs y los Bi-Fan ($X \rightarrow Y$, $X \rightarrow Z$, $X' \rightarrow Y$, $X' \rightarrow Z$), lo cual contribuyó a afianzar la hipótesis de que los motivos constituirían la base de procesamiento de información en el seno de los módulos funcionales [19].

Esquema de la tesis

Esta tesis presenta los resultados obtenidos en tres trabajos que versan sobre la autorregulación y los motivos de red: el primero pone el foco en un módulo funcional concreto, el sistema SOS de *E. coli*; el segundo estudio abarca la red transcripcional completa de esta misma bacteria; y el tercero, la práctica totalidad de los organismos procariotas cuyos genomas han sido secuenciados. Tratamos pues con tres estudios realizados a escalas o niveles biológicos de organización de muy distinto orden de magnitud.

Pequeña escala: Autorregulación y optimización del sistema SOS.

Tras un capítulo introductorio que trata de las propiedades dinámicas básicas de la autorregulación (sobre todo, de la negativa), el segundo capítulo de la tesis expone nuestros resultados teóricos y experimentales sobre el modo en el que la autorregulación optimiza la funcionalidad de un caso de motivo de red, el sistema SOS [20], del que acabamos de decir que es claramente asimilable a un módulo funcional. Este trabajo combina la modelización matemática con técnicas de biología sintética y con medidas de alta resolución temporal de la fluorescencia y la densidad óptica de cultivos bacterianos en crecimiento. La optimización se nos revela al evaluar una serie de propiedades tanto en el sistema natural como en su alternativa sintética en la que hemos eliminado la autorregulación del regulador maestro del SOS. Las propiedades evaluadas fueron las siguientes: i) la estabilidad del estado sin inducir (sin daño en el ADN), ii) el tiempo de respuesta del sistema, iii) la proporcionalidad de la respuesta al daño infligido, y iv) las curvas de crecimiento de los cultivos. En conjunto, nuestros resultados sugieren que el control autógeno del SOS evolucionó como una estrategia para responder óptimamente a un amplio rango de valores de la señal de entrada (el daño en el ADN), a la vez que se minimizan los costes de la respuesta.

Media escala: Autorregulación y ensamblaje de motivos de red.

¿Es posible generalizar la estrategia usada con el SOS al resto de aquellos motivos de red en los que la autorregulación se encuentre integrada? ¿Llegaríamos de nuevo a la conclusión de que su presencia implica una optimización del motivo correspondiente? Un buen indicio de esto lo constituye el hecho de que en los FFLs –el motivo de red más estudiado con la excepción quizás de la autorregulación– se ha identificado una presencia singular de elementos Y autorregulados [10]. Pero una generalización de este tipo asumiría implícitamente que todos y cada uno de los motivos de red han sido objeto de selección adaptativa en virtud de las propiedades dinámicas que se les atribuye, asunción sobre la que no hay unanimidad, puesto que, como hemos ya referido, ni siquiera la hay sobre el conjunto de los motivos en general.

El tercer capítulo expone nuestra aportación al debate sobre el origen adaptativo o neutral de los motivos de red, haciendo un énfasis especial en el papel que la autorregulación pueda estar jugando en el ensamblaje de éstos [21]. En nuestra aproximación hemos considerado que los argumentos que hay tras cada posición podrían conciliarse (al menos parcialmente) en el hecho de que en la dicotomía adaptativo/neutral el escenario correcto puede depender de cada motivo de red en particular. Y es que la señal estadística que dio pie a la descripción original de los motivos era un señal global basada en la comparación del recuento de estos motivos en la red natural frente a redes aleatorias alternativas derivadas de un modelo nulo [9], sin que se entrara en la relevancia estadística de cada uno de los motivos por separado. Nuestro análisis cubre este hueco mostrando efectivamente que la señal global de los motivos esconde señales individuales de muy distinta significatividad y que, por tanto, no todos los miembros de una clase de motivo dada son igual de improbables. No obstante, aun adoptando un punto de partida proclive a la selección neutral –en el que estos motivos eran catalogados como neutrales siempre y cuando concurriese alguna circunstancia que pudiera haberlos originado como un, por así decirlo, efecto colateral–, existe un número no desdeñable de motivos que se resisten a una explicación neutral. Curiosamente, como sucede con nuestro estudio del SOS, los experimentos que han contribuido a la visión de los motivos como entidades funcionales han sido realizados sobre motivos particulares que se que se encuentran entre aquéllos sobre los que menos duda cabe de su origen adaptativo [20, 22–24].

Gran escala: La autorregulación en la búsqueda de un código de reconocimiento proteína/ADN en la familia de factores de transcripción hélice-giro-hélice/LacI.

En el estudio anterior, una de las circunstancias que consideramos que podría llevar a un ensamblaje neutral de los motivos de red fue la arquitectura genómica, en concreto, la frecuente localización adyacente del operón que codifica a un factor de transcripción (TF) y alguno de los operones regulados por éste [25–28]. La regulación adyacente se da mucho más entre los TFs muy específicos, esto es, los que regulan a muy pocos operones; además, está muchas veces asociada a la propia autorregulación a través de la disposición divergente de los operones regulador y regulado, aunque también es frecuente la regulación adyacente en la que los dos operones tienen la misma orientación (como en el paradigmático caso del operón que codifica LacI y el operón regulado *lacZYA*).

Todo esto y la profusión de la autorregulación en la red de transcripción de *E. coli* –el 56 % de sus TFs está autorregulado [21]–, animaba a proponer una estrategia de búsqueda local de las secuencias de nucleótidos –o sitios de unión al ADN (*binding sites*, BS)– que son reconocidas por los TFs específicos. Y decimos local porque, tal y como se expone en el cuarto capítulo, esta estrategia consiste

en la búsqueda de los BSs en las zonas intergénicas adyacentes a la que codifica el TF. En nuestro caso, el objetivo último de la localización de los BSs consistía en revelar la existencia de un posible código de reconocimiento proteína/ADN que, dentro de una familia concreta de reguladores, relacionara los aminoácidos del TF y los nucleótidos del BS que establecen contactos específicos.

Resolver un código de la máxima universalidad exige que la búsqueda de los BSs se extienda sobre toda la filogenia procariota cuya información genómica esté disponible. A este respecto, el método basado en la regulación adyacente tiene la ventaja inmediata de no depender de la identificación de ortólogos entre los genes regulados –lo cual resulta problemático entre genomas distantes [29]– como sucede con los métodos en uso y es, por tanto, directamente aplicable a todos los genomas procariotas secuenciados a día de hoy. En concreto, nosotros lo hemos particularizado en la familia de TFs de LacI, cuyo dominio hélice-giro-hélice (*helix-turn-helix*, HTH) es el responsable de la unión al BS. Por su importancia histórica en el desarrollo del modelo de operón, la proteína LacI ha sido profusamente estudiada a nivel estructural, incluyendo su unión al ADN [30–34]. El modelo aceptado –y soportado por trabajos experimentales realizados con variantes mutados de LacI [35, 36]– hace recaer gran parte de la especificidad en tres de los seis aminoácidos de la hélice de reconocimiento –la segunda del dominio HTH, leído desde el extremo N-terminal [37].

Aunque el resto de los aminoácidos del HTH juega sobre todo un papel de soporte estructural, también condiciona la manera en la que se establecen los enlaces desde los aminoácidos que entran en contacto directo con el ADN. Esta dependencia del contexto estructural hace que la interacción aminoácido/nucleótido sea un fenómeno tridimensional complejo que aboca al fracaso todo intento de encontrar un código de reconocimiento proteína/ADN de carácter universal, es decir, un código que rija sobre todas las familias de TFs [38]. No obstante, cuando nos limitamos a un conjunto de proteínas en las que los aminoácidos que contactan el ADN están embebidos en un contexto estructural semejante, se puede intentar la búsqueda de soluciones a dicho código válidas dentro de este conjunto [39]; y la familia HTH-LacI ofrecía una oportunidad de encontrar tales soluciones que, si bien parciales, serían susceptibles de aplicación a un importante subconjunto de los TFs procariotas. Efectivamente, si bien el contexto estructural puede llegar a variar lo suficiente como para dar lugar a que TFs portadores de una misma secuencia de aminoácidos de reconocimiento tengan afinidades excluyentes por distintos BSs, nosotros hemos encontrado un código de referencia de gran cobertura dentro de esta familia.

A diferencia de los dos estudios anteriores, ya publicados [20, 21] y adjuntados como apéndices de la tesis, este último –realizado en colaboración con el profesor Eric Alm (Departamento de Ingeniería Civil y Medioambiental, Instituto de Tecnología de Massachusetts)– se encuentra en sus últimas fases de revisión.

Capítulo 1

Dinámica de la autorregulación

Uno de los objetos de estudio de la Biología de Sistemas es la dinámica de los mecanismos de control celulares basados en procesos de realimentación. Quizás el caso más sencillo de éstos sea la autorregulación o control autógeno de la transcripción, que es, ciñéndonos al ámbito procariota, la regulación de la transcripción de un operón por un factor de transcripción (TF) codificado en uno de los genes de dicho operón. A semejanza de lo que sucede en los sistemas no biológicos con realimentación [40], la autorregulación transcripcional –sobre todo la basada en un represor– se asocia, por vía de sus propiedades dinámicas, a una optimización de las funciones celulares [7, 8]. Las ventajas funcionales asociadas al control autógeno implicarían una fuerte selección adaptativa –en la red de transcripción que se conoce mejor, la de la bacteria *Escherichia coli*, más de la mitad de los TFs están autorregulados [9, 21, 41, 42].

En este capítulo vamos a introducir tales propiedades haciendo uso del sencillo circuito genético que aparece en la Figura 1.1. El circuito consta de dos genes que codifican sendos TFs: uno de ellos se une a la zona intergénica que antecede al gen que lo codifica[1]. Decimos entonces que este gen presenta autorregulación transcripcional. Además, existe una regulación transcripcional adicional ejercida por el segundo TF. Nos referiremos a este tipo de regulación como regulación *externa*.

Así pues, tratamos con un circuito con relaciones eminentemente transcripcionales en el que no existe más regulación postraduccional que la propia degradación de las proteínas –sea una degradación activa por proteasas o pasiva por crecimiento celular (más adelante ahondaremos en esto). Sin embargo, pese a su sencillez, nos bastará para introducir las propiedades fundamentales de la autorregulación. Para ello, en primer lugar –y partiendo de la cinética química de las especies moleculares involucradas– deduciremos las ecuaciones deterministas que rigen la dinámica del circuito. La linealización posterior de estas ecuaciones permitirá hacer uso de formalismos de transmisión de señal y ruido que fueron desarrollados

[1] Por simplicidad, consideramos operones de un solo gen.

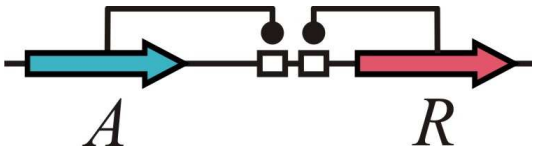


Figura 1.1: *Circuito autorregulatorio con regulación adicional externa. La transcripción del gen en rojo es regulada de manera autógena por el propio factor de transcripción R que este gen codifica. A esta autorregulación le acompaña la regulación adicional de carácter* externo *llevada a cabo por la proteína A, codificada en el gen en azul. Las cajas blancas representan los BSs.*

originalmente en disciplinas no biológicas, como las telecomunicaciones [43] o la teoría de control [44, 45].

Finalmente, a través de estos formalismos llegaremos a deducir de manera analítica las propiedades del sistema autorregulado. En particular, veremos como, frente a los sistemas carentes de control autógeno, la autorregulación negativa (esto es, la basada en un represor) está asociada a una mayor velocidad de respuesta ante cambios en la concentración de las proteínas reguladoras, y también al amortiguamiento de las fluctuaciones en la concentración. El tratamiento determinista (y en variable continua) de nuestro sistema requerirá alguna matización cuando abordemos el tratamiento de estas fluctuaciones, las cuales derivan de la verdadera naturaleza discreta y probabilística del proceso de producción y degradación de las proteínas.

1.1. De la química al modelo determinista

Una exposición general de cómo deducir un modelo determinista a partir de la cinética química puede encontrarse en [46]. Y de la aplicación al ámbito de los circuitos genéticos con realimentación son buenos ejemplos [47], [48] y [49]; en particular, nosotros vamos a seguir el tratamiento detallado que se da a este tipo de deducciones en el primero de estos tres trabajos.

Para empezar, volvamos a la Figura 1.1. Vamos a llamar R al monómero de la proteína que autorregula su transcripción y A al monómero de su regulador externo. La regulación no la llevan a cabo estos monómeros directamente, sino sus multímeros respectivos R_m y A_n. Vamos a asumir que existe un único sitio de unión (BS, por *binding site*) para cada uno estos multímeros en la zona operadora de R. Por tanto, existen cuatros estados posibles de ocupación de esta zona: S_0, con el operador sin ocupar por ningún TF; S_1, ocupado por el multímero A_m; S_2, ocupado por el multímero R_n; y S_3 ocupado a la vez por A_m y R_n. Al no sufrir A regulación transcripcional de ningún tipo, su región promotora siempre se encuentra en el estado constitutivo S_0'.

En cualquier sistema genético van a existir procesos que ocurren con tiempos característicos muy distinto. Ello permite que en muchas ocasiones se pueda asumir que las reacciones que se producen de manera más rápida están en equilibrio o cuasi-equilibrio [46]. Así pues, se consideran rápidas las reacciones de unión de los TFs al ADN,

$$\begin{aligned} S_0 + A_m &\underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} S_1\,, \\ S_0 + R_n &\underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} S_2\,, \\ S_1 + R_n &\underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} S_3\,, \\ S_2 + A_m &\underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} S_3\,; \end{aligned}$$

y los procesos de multimerización,

$$\begin{aligned} nR &\underset{k_{-5}}{\overset{k_5}{\rightleftharpoons}} R_n\,, \\ mA &\underset{k_{-6}}{\overset{k_6}{\rightleftharpoons}} A_m\,. \end{aligned} \tag{1.1}$$

Las constantes k's son las tasas de reacción directa (sobre las flechas) o inversa (debajo). En comparación con estas reacciones, se consideran lentos los procesos de transcripción, traducción y degradación [47]. Así, para el caso de R tenemos:

$$\begin{aligned} S_0 &\xrightarrow{\beta_R} S_0 + m_R, \\ S_1 &\xrightarrow{\rho_1\beta_R} S_1 + m_R, \\ S_2 &\xrightarrow{\rho_2\beta_R} S_2 + m_R, \\ S_3 &\xrightarrow{\rho_3\beta_R} S_3 + m_R, \\ m_R &\xrightarrow{\gamma_R} m_R + R, \\ m_R &\xrightarrow{\alpha_R} \emptyset, \\ R &\xrightarrow{\xi_R} \emptyset. \end{aligned} \tag{1.2}$$

donde β_R es la tasa de transcripción en ausencia de regulación (estado constitutivo) y los $\rho' s$ son números adimensionales que dan cuenta de la proporción que con ésta guardan las tasas de transcripción de los otros tres estados del promotor; m_R es el ARN mensajero (mARN) de R y γ_R la tasa con la que este mensajero se traduce; y α_R y ξ_R son las tasas de degradación de m_R y R, respectivamente.

Como el sitio operador de A sólo tiene accesible el estado constitutivo, el número de reacciones involucradas es menor:

$$\begin{aligned} S_0' &\xrightarrow{\beta_A} S_0' + m_A, \\ m_A &\xrightarrow{\gamma_A} m_A + A, \\ m_A &\xrightarrow{\alpha_A} \emptyset, \\ A &\xrightarrow{\xi_A} \emptyset. \end{aligned} \tag{1.3}$$

con los parámetros y variables definidos en correspondencia con los de R.

Vamos a asumir que las concentraciones son lo suficientemente grandes como para que éstas puedan representarse como variables continuas y, además, en todas las reacciones anteriores rija la *ley de acción de masas* [46, 50], es decir, que la probabilidad por unidad de tiempo de que se produzca una reacción venga dada por el producto de las concentraciones de los reactantes y la constante correspondiente[2]. Como estas constantes se definen por unidad de tiempo y de volumen, hay que multiplicar por este último, Ω, para obtener las tasas con las que se producen las reacciones. Finalmente, en la expresión que gobierna la tasa de acumulación de una especie molecular se han de tener en cuenta todas las reacciones en las que esta especie interviene (sea como reactante o como producto) y que conlleven una ganancia o pérdida en el número de moléculas de la misma. Por ejemplo, para el caso del número de promotores de R desocupados tendríamos:

$$\frac{\mathrm{d}S_o}{\mathrm{d}t} = -\Omega k_1 \frac{S_o}{\Omega}\frac{A_m}{\Omega} + \Omega k_{-1}\frac{S_1}{\Omega} - \Omega k_2 \frac{S_o}{\Omega}\frac{R_n}{\Omega} + \Omega k_{-2}\frac{S_2}{\Omega}. \tag{1.4}$$

De un modo análogo construiríamos las ecuaciones para el resto de especies moleculares. Antes de escribirlas notemos que, así formulada, las variables de la ecuación (1.4) expresan números absolutos de proteínas. Nosotros estamos interesados en describir el sistema en términos de concentraciones, de manera que el volumen, que en general depende del tiempo, quede integrado en las variables (ej. $[A] \equiv A/\Omega$). De momento, al dividir por $\Omega(t)$, el conjunto de ecuaciones del sistema queda como sigue:

[2]Constantes que lo serían siempre y cuando el sistema se encuentre en unas condiciones estables de, por ejemplo, temperatura o pH.

$$
\begin{aligned}
\frac{1}{\Omega}\frac{\mathrm{d}S_o}{\mathrm{d}t} &= -k_1[S_o][A_m] + k_{-1}[S_1] - k_2[S_o][R_n] + k_{-2}[S_2],\\
\frac{1}{\Omega}\frac{\mathrm{d}S_1}{\mathrm{d}t} &= k_1[S_o][A_m] - k_{-1}[S_1] - k_3[S_1][R_n] + k_{-3}[S_3],\\
\frac{1}{\Omega}\frac{\mathrm{d}S_2}{\mathrm{d}t} &= k_2[S_o][R_n] - k_{-2}[S_2] - k_4[S_2][A_m] + k_{-4}[S_3],\\
\frac{1}{\Omega}\frac{\mathrm{d}S_3}{\mathrm{d}t} &= k_3[S_1][R_n] - k_{-3}[S_3] + k_4[S_2][A_m] - k_{-4}[S_3],\\
\frac{1}{\Omega}\frac{\mathrm{d}R_n}{\mathrm{d}t} &= -k_2[S_o][R_n] + k_{-2}[S_2] - k_3[S_1][R_n] + k_{-3}[S_3] + k_5[R]^n - k_{-5}[R_n],\\
\frac{1}{\Omega}\frac{\mathrm{d}A_m}{\mathrm{d}t} &= -k_1[S_o][A_m] + k_{-1}[S_1] - k_4[S_2][A_m] + k_{-4}[S_3] + k_6[A]^m - k_{-6}[A_m],\\
\frac{1}{\Omega}\frac{\mathrm{d}R}{\mathrm{d}t} &= -k_5 n[R]^n + k_{-5} n[R_n] + \gamma_R[m_R] - \xi_R[R],\\
\frac{1}{\Omega}\frac{\mathrm{d}A}{\mathrm{d}t} &= -k_6 m[A]^m + k_{-6} m[A_m] + \gamma_A[m_A] - \xi_A[A],\\
\frac{1}{\Omega}\frac{\mathrm{d}m_R}{\mathrm{d}t} &= \beta_R([S_o] + \rho_1[S_1] + \rho_2[S_2] + \rho_3[S_3]) - \alpha_R[m_R],\\
\frac{1}{\Omega}\frac{\mathrm{d}m_A}{\mathrm{d}t} &= \beta_A[S'_o] - \alpha_A[m_A],\\
\frac{1}{\Omega}\frac{\mathrm{d}S'_o}{\mathrm{d}t} &= 0.
\end{aligned}
$$

La última ecuación es nula debido a que sólo consideramos un estado (constitutivo) para el promotor de A. Como las ecuaciones rápidas se consideran en equilibrio, los componentes con $k_{\pm i} (i = 1, 2, \ldots, 6)$ se van a cancelar mutuamente. Estas cancelaciones, de las que derivaremos un poco más adelante las constantes de equilibrio de las reacciones correspondientes, por lo pronto implican que las seis primeras ecuaciones del sistema anterior se anulan, quedando así éste reducido a uno de cuatro variables:

$$
\begin{aligned}
\frac{1}{\Omega}\frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma_R[m_R] - \xi_R[R],\\
\frac{1}{\Omega}\frac{\mathrm{d}A}{\mathrm{d}t} &= \gamma_A[m_A] - \xi_A[A],\\
\frac{1}{\Omega}\frac{\mathrm{d}m_R}{\mathrm{d}t} &= \beta_R([S_o] + \rho_1[S_1] + \rho_2[S_2] + \rho_3[S_3]) - \alpha_R[m_R],\\
\frac{1}{\Omega}\frac{\mathrm{d}m_A}{\mathrm{d}t} &= \beta_A[S'_o] - \alpha_A[m_A]. \qquad (1.5)
\end{aligned}
$$

En estas cuatro ecuaciones los términos de la izquierda aún están expresados en cantidades absolutas. Ahora bien, aplicando la regla de la cadena a la derivada de

la concentración de R se obtiene que el término de la izquierda de la primera de ellas puede reescribirse como:

$$\frac{1}{\Omega}\frac{\mathrm{d}R}{\mathrm{d}t} = \frac{\mathrm{d}[R]}{\mathrm{d}t} + \frac{1}{\Omega}\frac{\mathrm{d}\Omega}{\mathrm{d}t}[R]. \tag{1.6}$$

Y lo mismo para el caso de A, m_R y m_A. El segundo término de la derecha supondrá la aparición en las ecuaciones diferenciales de un componente de degradación $\frac{1}{\Omega}\frac{\mathrm{d}\Omega}{\mathrm{d}t}$ asociado al crecimiento y que, independientemente de la especie molecular de que se trate, contribuye a la disminución de la concentración no por vía de una degradación molecular propiamente dicha (tipo ξ_R, ξ_A, α_R o α_A), sino por el efecto de dilución en un volumen creciente. Así, con todo lo anterior, obtenemos las siguientes ecuaciones en las que las variables están ya siempre expresadas en concentraciones:

$$\begin{aligned}
\frac{\mathrm{d}[R]}{\mathrm{d}t} &= \gamma_R[m_R] - \left(\xi_R + \frac{1}{\Omega}\frac{\mathrm{d}\Omega}{\mathrm{d}t}\right)[R], \\
\frac{\mathrm{d}[A]}{\mathrm{d}t} &= \gamma_A[m_A] - \left(\xi_A + \frac{1}{\Omega}\frac{\mathrm{d}\Omega}{\mathrm{d}t}\right)[A], \\
\frac{\mathrm{d}[m_R]}{\mathrm{d}t} &= \beta_R([S_o] + \rho_1[S_1] + \rho_2[S_2] + \rho_3[S_3]) - \left(\alpha_R + \frac{1}{\Omega}\frac{\mathrm{d}\Omega}{\mathrm{d}t}\right)[m_R], \\
\frac{\mathrm{d}[m_A]}{\mathrm{d}t} &= \beta_A[S'_o] - \left(\alpha_A + \frac{1}{\Omega}\frac{\mathrm{d}\Omega}{\mathrm{d}t}\right)[m_A].
\end{aligned} \tag{1.7}$$

El término de degradación $\frac{1}{\Omega}\frac{\mathrm{d}\Omega}{\mathrm{d}t}$ en su forma general no es una constante. Pero sí lo es si asumimos un modelo de crecimiento exponencial,

$$\frac{\mathrm{d}\Omega}{\mathrm{d}t} = \xi_c\,\Omega.$$

Efectivamente, en este caso el término de degradación es simplemente la constante ξ_c. Además, evaluando la expresión anterior ya integrada, $\Omega(t) = \Omega(t_o)\,e^{\xi_c(t-t_o)}$, a lo largo del tiempo de un ciclo celular t_c –en el que el volumen se duplica, $\Omega(t_o + t_c) = 2\Omega(t_o)$– se obtiene el valor de la constante de degradación asociada al crecimiento en función de la duración del ciclo celular: $\xi_c = \log 2/t_c$.

Para comparar el orden de magnitud de ξ_c frente a los términos de degradación convencionales, tengamos en cuenta que estos últimos son inversamente proporcionales a la vida media de la proteína respectiva, como podemos comprobar anulando el componente de producción de proteína en la primera de las ecuaciones del sistema (1.5). En ese caso, la solución de esta ecuación es un decaimiento exponencial en el número absoluto de proteínas existente en un tiempo dado t_o: $R(t) = R(t_o)e^{-\xi_R(t-t_o)}$. Si ahora definimos la semivida t_R de la proteína R como el tiempo necesario para que este decaimiento exponencial reduzca el número absoluto de proteínas a la mitad de su valor en un instante dado, $R(t_o + t_R) = \frac{1}{2}R(t_o)$,

se llega a la relación $\xi_R = \log 2/t_R$. Por supuesto, lo mismo aplica al resto de especies moleculares.

A partir de ahora vamos a unificar en un única término genérico las dos contribuciones a la degradación de cada una de las proteínas y mensajeros:

$$\begin{aligned}
\delta_R &\equiv \xi_c + \xi_R = \log 2 \left(\frac{1}{t_c} + \frac{1}{t_R} \right), \\
\delta_A &\equiv \xi_c + \xi_A = \log 2 \left(\frac{1}{t_c} + \frac{1}{t_A} \right), \\
\delta_{m_R} &\equiv \xi_c + \alpha_R = \log 2 \left(\frac{1}{t_c} + \frac{1}{t_{m_R}} \right), \\
\delta_{m_A} &\equiv \xi_c + \alpha_A = \log 2 \left(\frac{1}{t_c} + \frac{1}{t_{m_A}} \right).
\end{aligned}$$

Como se ve, la comparación entre las distintos componentes de la degradación se convierte en una comparación entre los órdenes de magnitud de la semivida y la duración de un ciclo celular –este último en torno a los 20 minutos en el caso de cultivos de *Escherichia coli* en medios ricos. Para proteínas muy estables –con semividas de horas– la degradación se suele aproximar por el componente debido al proceso de dilución por crecimiento, ξ_c. Las semividas de los mARN en *E. coli* suelen ser del orden de t_c [51].

Vamos a asumir que la dinámica de la transcripción y degradación de m_R y m_A es rápida en comparación con la correspondiente a las proteínas [47]. Así podemos hacer uso de la aproximación que considera que las concentraciones de los mensajeros están en un estado cuasiestacionario con respecto a las de las proteínas ($\frac{\mathrm{d}[m_R]}{\mathrm{d}t} \sim 0$, $\frac{\mathrm{d}[m_A]}{\mathrm{d}t} \sim 0$). Finalmente, esta aproximación permite despejar la concentración de los mensajeros respectivos y reducir el sistema (1.7) a uno de dos variables ($[R]$ y $[A]$):

$$\begin{aligned}
\frac{\mathrm{d}[R]}{\mathrm{d}t} &= \gamma_R \frac{\beta_R}{\delta_{m_R}} ([S_o] + \rho_1[S_1] + \rho_2[S_2] + \rho_3[S_3]) - \delta_R[R], \\
\frac{\mathrm{d}[A]}{\mathrm{d}t} &= \gamma_A \frac{\beta_A}{\delta_{m_A}} [S'_o] - \delta_A[A].
\end{aligned}$$

Como habíamos anticipado, la consideración de que las reacciones de los TFs con el promotor y las de multimerización están en equilibrio permite definir las constantes de los equilibrios respectivos. Así, para la unión de los multímeros al

ADN, tenemos:

$$K_1 = \frac{[S_1]}{[S_0][A_m]} = \frac{k_1}{k_{-1}},$$
$$K_2 = \frac{[S_2]}{[S_0][R_n]} = \frac{k_2}{k_{-2}},$$
$$K_3 = \frac{[S_3]}{[S_1][R_n]} = \frac{k_3}{k_{-3}},$$
$$K_4 = \frac{[S_3]}{[S_2][A_m]} = \frac{k_4}{k_{-4}},$$

y para el proceso de multimerización:

$$K_R = \frac{[R_n]}{[R]^n} = \frac{k_5}{k_{-5}},$$
$$K_A = \frac{[A_m]}{[A]^m} = \frac{k_6}{k_{-6}}.$$

La concentración de un estado cualquiera del promotor S_i ($i = 0, 1, 2, 3$) es proporcional a la probabilidad P_{Si} de este estado [48]. Así, a partir de las expresiones para las constantes de equilibrio obtenemos la relación entre las probabilidades: $P_{S1}/P_{S0} = K_1[A_m]$, $P_{S2}/P_{S0} = K_2[R_n]$, $P_{S3}/P_{S1} = K_3[R_n]$, $P_{S3}/P_{S2} = K_4[A_m]$, y de aquí directamente: $P_{S3}/P_{S0} = K_1K_3[A_m][R_n] = K_2K_4[A_m][R_n]$, lo que implica que

$$K_1K_3 = K_2K_4. \tag{1.8}$$

Esta relación se satisface de manera trivial si la unión/desunión de uno de los TFs al ADN no depende de la presencia o ausencia del otro TF. En ese caso $K_1 = K_4$ y $K_2 = K_3$. En el caso más general, en el que las proteínas se afectan entre sí, la relación (1.8) se cumple si $K_3 = \epsilon K_2$ y $K_4 = \epsilon K_1$. La constante ϵ es el *parámetro de cooperatividad*: si la unión de una de las especies aumenta la probabilidad de enlace de la otra, entonces $\epsilon > 1$; por el contrario, si la disminuye, tenemos que $0 \leq \epsilon < 1$.

Como el promotor ha de estar forzosamente en uno de los cuatro estados, la norma que lleva de concentraciones a probabilidades no es sino la suma de las concentraciones de cada uno de los estados. Dicho de otro modo, esta norma no es sino la concentración del promotor cualquiera que sea su estado de ocupación: $[S_R] \equiv [S_0] + [S_1] + [S_2] + [S_3]$ para el represor y $[S_A] \equiv [S_0']$ para el activador. Con todo lo anterior, el aspecto que adquiere nuestro sistema de ecuaciones es el que sigue:

$$\frac{\mathrm{d}[A]}{\mathrm{d}t} = \gamma_A \frac{\beta_A}{\delta_{m_A}}[S_A] - \delta_A[A],$$
$$\frac{\mathrm{d}[R]}{\mathrm{d}t} = f([A],[R]) - \delta_R[R],$$

con

$$f([A],[R]) = \gamma_R \frac{\beta_R}{\delta_{m_R}}[S_R] \frac{1+\rho_1 K_1 K_A [A]^m + \rho_2 K_2 K_R [R]^n + \rho_3 \epsilon K_1 K_A [A]^m K_2 K_R [R]^n}{1+K_1 K_A [A]^m + K_2 K_R [R]^n + \epsilon K_1 K_A [A]^m K_2 K_R [R]^n}.$$

El modelo de crecimiento exponencial también se puede aprovechar para aproximar mediante una solución analítica sencilla la concentración de promotores $[S_i]$ $(i = A, R)$ por su valor promedio en el ciclo celular. Así, si N_i $(i = A, R)$ es el número de copias del gen que codifica al regulador antes de la replicación del cromosoma en t_r, donde este número se duplica, entonces:

$$[S_i(t)] \sim <[S_i(t)]> = \frac{N_i}{t_c}\left[\int_0^{t_r} \mathrm{d}t \frac{1}{\Omega(t)} + \int_{t_r}^{t_c} \mathrm{d}t \frac{2}{\Omega(t)}\right] = \frac{N_i}{\Omega_o\ 2^{t_r/t_c}\ \log 2} \quad i = A, R.$$

Esta aproximación tiene validez cuando tratamos con poblaciones de células, puesto que, con el astronómico número de bacterias involucrado en una medida poblacional de laboratorio, la asincronía de los ciclos celulares implica un probabilidad uniforme de encontrar a una bacteria dada en cualquier momento del ciclo celular. Por el contrario, en el caso de medidas con unas pocas células habría que considerar explícitamente la duplicación del promotor en cierto momento del ciclo celular y la división de la bacteria.

La notación se simplifica agrupando las constantes que se multiplican. Por un lado, en el concepto de *actividad del promotor* (AP) agrupamos al conjunto $\Pi_i \equiv \gamma_i \frac{\beta_i}{\delta_{m_i}} < [S_i(t)] > (i = A, R)$; y por otro, las constantes de equilibrio para los procesos de unión al ADN y multimerización se integran en las constantes de disociación[3]: $k_A \equiv (K_1 K_A)^{-m}$ y $k_R \equiv (K_2 K_R)^{-n}$. A partir de ahora, para simplificar la notación, vamos a suprimir los corchetes que denotan concentraciones, con lo que la forma final de nuestro modelo determinista bidimensional es la que sigue:

$$\begin{aligned} \frac{\mathrm{d}A}{\mathrm{d}t} &= \Pi_A - \delta_A A, \\ \frac{\mathrm{d}R}{\mathrm{d}t} &= f(A,R) - \delta_R R, \end{aligned} \tag{1.9}$$

siendo

$$f(A,R) = \Pi_R \frac{1+\rho_1 \left(\frac{A}{k_A}\right)^m + \rho_2 \left(\frac{R}{k_R}\right)^n + \rho_3\ \epsilon \left(\frac{A}{k_A}\right)^m \left(\frac{R}{k_R}\right)^n}{1+\left(\frac{A}{k_A}\right)^m + \left(\frac{R}{k_R}\right)^n + \epsilon \left(\frac{A}{k_A}\right)^m \left(\frac{R}{k_R}\right)^n}.$$

[3]Como su nombre indica, cuanto mayores sean estas constantes menores serán las afinidades de los TFs por la zona reguladora.

El lado derecho de las ecuaciones diferenciales (1.9) consta de un término de producción de proteínas (o AP) y un término de degradación. Nótese que las constantes Π_A y Π_R representan la AP en ausencia de regulación; circunstancia que se da siempre en el caso de A, proteína no sujeta a regulación transcripcional. En cambio, la AP de R es toda la función f, en la que Π_R está modulada por un término adimensional que recoge la influencia de los reguladores. La función $f(A, R)$ es la responsable de la no linealidad del sistema.

Podemos reencontrarnos con el origen probabilístico de nuestro modelo determinista si interpretamos f como el promedio de la AP sobre una distribución de probabilidad discreta con cuatro estados accesibles –los distintos estados de ocupación del promotor, S_i $(i = 0, 1, 2, 3)$. Los niveles de actividad asociados a estos estados en unidades de Π_R son 1, ρ_1, ρ_2 y ρ_3, respectivamente; y las probabilidades normalizadas de los estados vendrían dadas por

$$\begin{aligned} P_{S0} &= Z^{-1}, \\ P_{S1} &= Z^{-1}\left(\frac{A}{k_A}\right)^m, \\ P_{S2} &= Z^{-1}\left(\frac{R}{k_R}\right)^n, \\ P_{S3} &= Z^{-1}\epsilon\left(\frac{A}{k_A}\right)^m\left(\frac{R}{k_R}\right)^n, \end{aligned} \tag{1.10}$$

siendo Z la función de partición:

$$Z = 1 + \left(\frac{A}{k_A}\right)^m + \left(\frac{R}{k_R}\right)^n + \epsilon\left(\frac{A}{k_A}\right)^m\left(\frac{R}{k_R}\right)^n.$$

Las probabilidades dependen de las concentraciones de los TFs, las constantes de disociación y el parámetro de cooperatividad ϵ. La concentración de cada regulador aparece dividida por la constante de disociación correspondiente, la cual tiene también unidades de concentración; por tanto, la probabilidad de que un TF se una al promotor será tanto mayor cuanto mayor sea la concentración del mismo y menor su constante de disociación[4].

Las dos ecuaciones el sistema (1.9) están desacopladas, puesto que de la primera de ellas podemos obtener directamente la solución estacionaria para la concentración de A:

$$A_* = \frac{\Pi_A}{\delta_A},$$

[4]A la hora de mapear el comportamiento del sistema bajo distintos estados del promotor es preferible manipular las constantes de afinidad, puesto que son parámetros y no variables del sistema. Esto se puede extender a todas las definiciones que se hagan en relación a configuraciones particulares del sistema. Por ejemplo, la ausencia de regulación la vincularemos a unos valores infinitos de las constantes de disociación en lugar de a concentraciones nulas de los reguladores.

que es una constante debido a que la AP, en este caso, no varía. Nuestro problema queda así reducido a un sistema de una sola variable cuya solución estacionaria R_* (Figura 1.5.A) está determinada por la función implícita siguiente:

$$R_* = \frac{1}{\delta_R}\, f(A_*, R_*).$$

En ausencia de regulación recuperamos una expresión para la solución estacionaria de R semejante a la de A,

$$R_o \equiv R_*|_{k_A\to\infty, k_R\to\infty} = \frac{\Pi_R}{\delta_R}.$$

Los extremos de R_* se alcanzan cuando dominan estados asociados a la máxima o mínima AP,

$$\begin{aligned} max(R_*/R_o) &= max\{1, \rho_1, \rho_2, \rho_3\}, \\ min(R_*/R_o) &= min\{1, \rho_1, \rho_2, \rho_3\}, \end{aligned} \tag{1.11}$$

lo cual se puede forzar matemáticamente manipulando los valores de las constantes de disociación y cooperatividad; por ejemplo, si máx$\{1, \rho_1, \rho_2, \rho_3\} = \rho_1$ se alcanzará el máximo de actividad si $k_A \to 0$.

1.2. Aproximación lineal

La linealización del sistema de ecuaciones (1.9) en la vecindad del estado estacionario (A_*, R_*) supone una buena aproximación al modelado de la dinámica del sistema genético si las concentraciones no se apartan demasiado de este estado. En particular, esto suele ser así cuando los cambios en las concentraciones obedecen sólo a fluctuaciones aleatorias (o ruido) en torno al estado estacionario. Además, la linealización permite llegar a soluciones analíticas que describen las propiedades de transmisión del sistema. Estas soluciones, que son un lugar común de la teoría de control de sistemas lineales [43–45], están expuestas de manera genérica y sucinta en el Apéndice A; en esta sección las particularizaremos en el circuito genético que nos ocupa.

Así, sean $x(t)$ y $y(t)$ pequeños desplazamientos de A y R con respecto a las concentraciones estacionarias A_* y R_*, respectivamente:

$$\begin{aligned} A(t) &= A_* + x(t), \\ R(t) &= R_* + y(t). \end{aligned}$$

Si se sustituyen estas expresiones en (1.9) y se retienen sólo los términos lineales del desarrollo de Taylor de la función f, se obtienen las ecuaciones lineales que

rigen el comportamiento del sistema en el entorno de $(A_*,\, R_*)$:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= -\delta_A\, x, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= b\, x(t) - (\delta_R + a)\, y(t), \end{aligned} \tag{1.12}$$

con

$$b \equiv \left.\frac{\partial f}{\partial A}\right|_{A_*,R_*} \qquad a \equiv -\left.\frac{\partial f}{\partial R}\right|_{A_*,R_*}.$$

Teniendo en cuenta la forma de las probabilidades P_{Si} $(i = 0, 1, 2, 3)$ de los distintos estados del operador en (1.10), podemos obtener con facilidad la expresión que en nuestro sistema genético tienen los parámetros lineales a y b,

$$b = m\,\delta_R\,\frac{R_*}{A_*}\left[\left(\frac{\rho_1}{R_*/R_o} - 1\right) P_{S1} + \left(\frac{\rho_3}{R_*/R_o} - 1\right) P_{S3}\right], \tag{1.13}$$

$$a = -n\,\delta_R\left[\left(\frac{\rho_2}{R_*/R_o} - 1\right) P_{S2} + \left(\frac{\rho_3}{R_*/R_o} - 1\right) P_{S3}\right]. \tag{1.14}$$

Nótese que b depende de los estados del promotor en los que hay un multímero de A unido a éste (con independencia de que lo haya o no de R) y que a depende de aquéllos con presencia de un multímero de R.

1.2.1. Definiciones absolutas y relativas de activación y represión

Consideremos cualquiera de los tres estados del promotor en los que hay algún TF unido a la zona operadora, S_i $(i = 1, 2$ ó $3)$. Habitualmente nos referimos a este estado como de activación (represión) si se produce un incremento (disminución) de la tasa de transcripción con respecto al estado S_o, en el que el promotor está sin ocupar. Esto matemáticamente se traduce en unas condiciones bien sencillas y sin ambigüedad: $\rho_i > 1$ implica activación mientras que para $\rho_i < 1$ tenemos represión[5].

Las definiciones anteriores son *absolutas* en el sentido de que están referidas a un concepto invariante: la AP en el estado S_o, Π_R; por tanto, el carácter de estado de activación o represión es también invariante. Pero, supongamos ahora,

[5]Obsérvese que reservamos los calificativos de activador y represor para los estados y no para las proteínas. Si sólo interviniesen estados con un sólo tipo de TF unido a la zona operadora (es decir, si $\epsilon = 0$) podríamos extender el calificativo al propio factor implicado, pero ello no es siempre posible cuando hay estados en los que el promotor está ocupado por distintas especies de reguladores a la vez. Por ejemplo, si $\rho_1 > 1$, $\rho_2 < 1$ y $\rho_3 < 1$, ¿cómo calificamos a la proteína A, que interviene en los estados S_1 y S_3?

para simplificar la discusión que sigue, que ambos TFs compiten por una misma zona del ADN[6]. Con esto, el sistema de ecuaciones (1.9) se reduce al siguiente:

$$\begin{aligned}\frac{\mathrm{d}A}{\mathrm{d}t} &= \Pi_A - \delta_A A,\\ \frac{\mathrm{d}R}{\mathrm{d}t} &= \Pi_R \frac{1+\rho_1\left(\frac{A}{k_A}\right)^m+\rho_2\left(\frac{R}{k_R}\right)^n}{1+\left(\frac{A}{k_A}\right)^m+\left(\frac{R}{k_R}\right)^n} - \delta_R R;\end{aligned} \tag{1.15}$$

las expresiones de a y b también se simplifican por la imposibilidad del estado S_3:

$$b = m\,\delta_R \frac{R_*}{A_*}\left(\frac{\rho_1}{R_*/R_o}-1\right)P_{S1}, \tag{1.16}$$

$$a = -n\,\delta_R\left(\frac{\rho_2}{R_*/R_o}-1\right)P_{S2}. \tag{1.17}$$

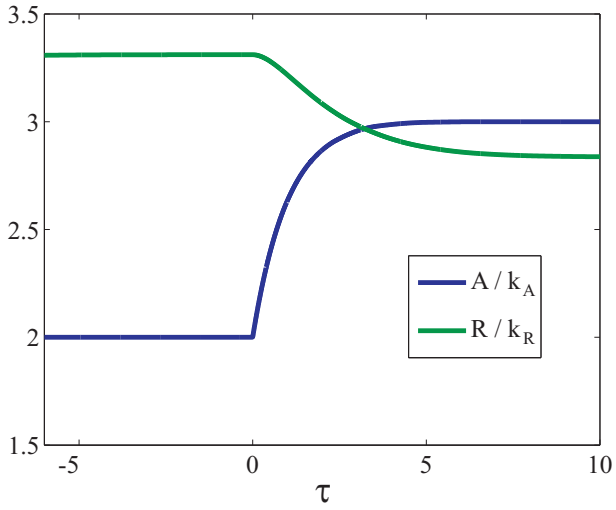


Figura 1.2: *Represión efectiva por activador. En $t = 0$ la AP de A crece súbitamente de $\tilde{A}_* = 2$ a $\tilde{A}_* = 3$. Ello va acompañado de un crecimiento en la concentración de A (curva azul) y una caída en la de R (en verde). Parám.: $m = n = 2$, $\rho_1 = 2$, $\rho_2 = 4$, $\tilde{R}_o = 1$, $\delta_R = \delta_A$. Parámetros y variables adimensionales definidos en la nota 7.*

Supongamos además que tanto A como R son activadores según la definición anterior, pero A lo es con menos efectividad, $1 < \rho_2 < \rho_1$. La Figura 1.2 muestra

[6]Es decir, que el estado S_3 con los dos reguladores unidos al ADN a la vez no se puede dar debido a que la unión de uno de los reguladores al promotor excluye la posibilidad de unión del otro ($\epsilon = 0$). Ello nos permite hablar en este caso en función de proteínas o de estados a nuestra conveniencia (véase la nota anterior).

(en variables adimensionales)[7] la dinámica del sistema (1.15), el cual, partiendo de una situación estacionaria (A_*, R_*), sufre el súbito incremento de la AP de A en $t = 0$. En este caso, el incremento de la concentración de A aumenta la probabilidad del estado S_1 en el que la activación es menos efectiva, lo que lleva a una caída de la concentración de R. Nótese que el valor de ρ_1 es inferior al estado estacionario (normalizado) de partida, R_*/R_o[8]. Esto hace que la expresión entre paréntesis en (1.16) sea negativa y, con ello, el valor de la función de transferencia de entrada b –lo que implica que la concentración de R se va a mover en sentido opuesto a la de A. En conclusión, bajo estas circunstancias el activador A se estaría comportando *de manera efectiva* como un represor.

Efectos como éste aconsejarían definir la activación y la represión en relación al estado estacionario R_*: un estado S_i es de activación efectiva si $\rho_i > R_*/R_o$, y al contrario para el caso de la represión. Estas son definiciones *relativas* pues, a diferencia de lo que sucede con las anteriores, todo estado cuyo ρ no marque ni el máximo ni el mínimo posibles –véase (1.11)– para R_*/R_o, puede funcionar tanto como activador o represor efectivo, dependiendo de la concentración estacionaria en la que esté instalada el sistema.

La importancia de todo esto se revela a la hora de establecer los signos de las expresiones (1.13) y (1.14), pues para cada uno de los paréntesis que aparece en estas expresiones, $\frac{\rho_i}{R_*/R_o} - 1$ $(i = 1, 2, 3)$, su signo será positivo o negativo si el estado es, bajo una definición relativa, de activación o represión, respectivamente.

---

[7] Las ecuaciones (1.15) se pueden reescribir en términos de variables adimensionales,

$$\begin{aligned} \frac{\delta_R}{\delta_A}\frac{\mathrm{d}\tilde{A}}{\mathrm{d}\tau} &= \tilde{A}_* - \tilde{A}, \\ \frac{\mathrm{d}\tilde{R}}{\mathrm{d}\tau} &= \tilde{R}_o\,\frac{1+\rho_1\tilde{A}^m+\rho_2\tilde{R}^n}{1+\tilde{A}^m+\tilde{R}^n} - \tilde{R}, \end{aligned}$$

donde $\tau \equiv t\,\delta_R$, $\tilde{A} \equiv A/k_A$, $\tilde{R} \equiv R/k_R$ y, en consecuencia, $\tilde{A}_* = A_*/k_A = \frac{\Pi_A}{\delta_A\,k_A}$ y $\tilde{R}_o = R_o/k_R = \frac{\Pi_R}{\delta_R\,k_R}$. Nótese que la definición de τ equivale a medir el tiempo en unidades de un tiempo característico del sistema. La normalización de las concentraciones de los TFs por las constantes de afinidad respectivas permite valorar directamente dichas concentraciones en relación con los umbrales de la acción reguladora. La máxima sensibilidad de la respuesta del sistema a los cambios en la concentración de uno de sus reguladores se obtiene cuando la versión adimensional de esta concentración ronda valores en torno a la unidad (como sucede en nuestro ejemplo).

[8] Esto se puede ver directamente en la Figura 1.2, puesto que, como en este caso $\tilde{R}_o = 1$,

$$R_*/R_o = \tilde{R}_*/\tilde{R}_o = \tilde{R}_* \simeq 3.3,$$

que es mayor que $\rho_1 = 2$. Además, cuanto mayor sea la probabilidad del estado S_1, más se acercará la concentración estacionaria $\tilde{R}_*$ a ρ_1, pues, a partir de (1.15) tenemos que

$$\lim_{k_A \to 0} R_*/R_o = \rho_1.$$

1.2.2. Condiciones de autorregulación transcripcional negativa

¿Qué valores han de tener por tanto los parámetros ρ's del sistema general (1.9) para que en la linealización tengamos una autorregulación transcripcional negativa[9]? De la discusión del apartado anterior se deduce que una definición absoluta de S_2 y S_3 como estados de represión no garantizaría que a fuese positivo en todas las circunstancias pues el signo de a depende de cuál sea el signo de las expresiones entre paréntesis en (1.14). En general, se puede demostrar, partiendo de (1.14), que, para que a sea mayor que cero, se ha de satisfacer lo siguiente:

$$\begin{aligned} \frac{\rho_2}{R_*/R_o} &= 1+\eta, \\ \frac{\rho_3}{R_*/R_o} &< 1-\frac{\eta}{\epsilon\theta} \quad \text{con} \quad -1 \leq \eta < \epsilon\theta \quad \text{y} \quad \theta \equiv \left(\frac{A}{k_A}\right)^m, \end{aligned}$$

relaciones que, como se ve, no excluyen que uno de los estados S_2 y S_3 lo sea de activación efectiva. No obstante, en el caso particular de que $\rho_2 = \rho_3$, estas relaciones implican que tanto ρ_2 como ρ_3 han de ser menores que R_*/R_o (represión efectiva).

El estado estacionario R_* depende de, entre otros factores, cuáles sean los valores de las constantes de disociación k_A y k_R. Para garantizar que, con independencia del valor de estas constantes –es decir, para cualquier estado estacionario accesible al sistema–, el valor de a sea mayor que cero, la tasa de transcripción en S_2 y S_3 ha de ser la mínima de las cuatro asociadas a los correspondientes estados del promotor:

$$\rho_2 = \rho_3 < \text{mín}\{1, \rho_1\},$$

puesto que, como vimos en (1.11), el mínimo de R_*/R_o lo marca el mínimo del conjunto $\{1, \rho_1, \rho_2, \rho_3\}$. La condición anterior supone que el enlace del multímero de R a la zona operadora reprime la transcripción con independencia de que el otro TF esté unido o no al promotor. Vamos a considerar que este es el caso de nuestro sistema, con lo que a va a ser positivo en todo momento y, para cualquier valor de las concentraciones, los estados S_2 y S_3 son siempre de represión en términos absolutos ($\rho_2 = \rho_3 < 1$). Aunque no se excluye que S_1 pueda ser también un estado de represión, siempre que conllevara una represión más liviana ($\rho_2 = \rho_3 < \rho_1 < 1$), en general (y pese a lo dicho en la nota 9) supondremos que S_1 es un estado de activación ($\rho_1 > 1$), lo cual implica valores de b positivos.

[9] Es decir, para que a sea mayor que cero. El signo de b no nos preocupa en principio puesto que las propiedades de transmisión y filtrado del sistema no dependen de dicho signo; por la linealidad del sistema, si $y(t)$ es la respuesta del sistema a $x(t)$, $-y(t)$ lo es a la entrada $-x(t)$ (véase el Apéndice A).

1.2.3. Solución estacionaria del sistema lineal

En estas condiciones de control autógeno represivo la solución estacionaria (x_*, y_*) del sistema lineal (1.12) no es sino la anulación de las perturbaciones. Efectivamente, si en $t = 0$ situamos el sistema en una posición $(x_o,\ y_o)$, éste relajará a $(x_* = 0,\ y_* = 0)$ de forma exponencial. Explícitamente, la solución del sistema de ecuaciones con las condiciones iniciales dadas es

$$\begin{aligned} x(t) &= x_o e^{-\delta_A t}, \\ y(t) &= y_o e^{-(\delta_R + a)t} + \frac{b\, x_o}{\delta_R + a - \delta_A} \left(e^{-\delta_A t} - e^{-(\delta_R + a)t} \right), \end{aligned} \tag{1.18}$$

siempre y cuando $\delta_R + a \neq \delta_A$. En caso caso contrario, hay que sustituir la solución para $y(t)$ por:

$$y(t) = \left(y_o + \frac{b\, x_o\, t}{1 + \delta_A t - \delta_A t^2} \right) e^{-\delta_A t}.$$

La anulación de las perturbaciones se produce si $a > -\delta_R$ (los términos de degradación son siempre positivos); la consiguiente restauración del estado estacionario global $(A_*,\ R_*)$ de partida implica que este estado es estable. Esto significa que la condición de autorregulación *transcripcional* negativa *sensu stricto*, que requiere que el valor de a sea mayor que cero, puede relajarse un poco, puesto que (1.18) deja claro que para alcanzar la estabilidad basta con que esta condición la satisfaga la suma de las dos componentes de la realimentación, $a + \delta_R > 0$.

Lo anterior se debe a que, desde un punto de vista matemático, la degradación δ_R y la autorregulación transcripcional a juegan papeles análogos en (1.12). De todos modos, aunque en rigor la autorregulación tiene estas dos componentes, el término autorregulación suele ceñirse, en el ámbito de la Biología, a la componente transcripcional a. Por ello, es preferible hablar de *realimentación*, concepto proveniente de la Teoría de Control, cuando nos refiramos al conjunto de estas contribuciones. Podemos estimar la relación entre ellas haciendo $\rho_2 = 0$ (estado S_2 de represión total) en (1.17):

$$a = n\, \delta_R P_{S2},$$

con lo que el orden de magnitud de ambas contribuciones es, en general, comparable. Por tanto, no debe desdeñarse la contribución a la realimentación de la componente de degradación[10], la cual, además, se basta para que, en ausencia de autorregulación transcripcional negativa, un sistema genético pueda alcanzar un estado estacionario estable (véase el caso de la variable $x(t)$ en las soluciones anteriores).

[10]El término de degradación incluye, como hemos visto más arriba, el fenómeno de dilución por crecimiento.

1.3. Transmisión de señal y ruido

En el sistema que acabamos de resolver sólo hemos considerado una fluctuación puntual que, además, ha sido introducida de una manera un tanto artificiosa en forma de unas condiciones iniciales dadas. Sin embargo, en los sistemas naturales las concentraciones están fluctuando de manera continua e inevitable debido al carácter discreto y temporalmente estocástico de los procesos de producción y degradación de las proteínas. Este ruido, por provenir de la estocasticidad inherente de los procesos bioquímicos, se denomina *ruido intrínseco* [52]. Por ello, un modelado más realista del sistema que dé cuenta del ruido intrínseco exige la inclusión de sendas señales de fluctuación, $u(t)$ y $v(t)$, en las ecuaciones del sistema lineal (1.12), esto es,

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= u(t) - \delta_A\, x(t), \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= v(t) + b\, x(t) - (\delta_R + a)\, y(t). \end{aligned} \tag{1.19}$$

Por otro lado, para cualquier sistema genético natural existe un segundo tipo de ruido proveniente esta vez de la propagación a través del sistema de fluctuaciones en la concentración de componentes celulares *externos* al mismo (como las polimerasas y los TFs externos que lo regulan) y que por ello se denomina *ruido extrínseco* [52]. En nuestro sistema, aunque sencillo en su modelado, también vamos a encontrar este tipo de ruido sin necesidad de añadir términos adicionales a las ecuaciones: se trata del componente de ruido en la concentración de R que deriva del hecho de que su producción está regulada por el TF externo A, cuya concentración también fluctúa.

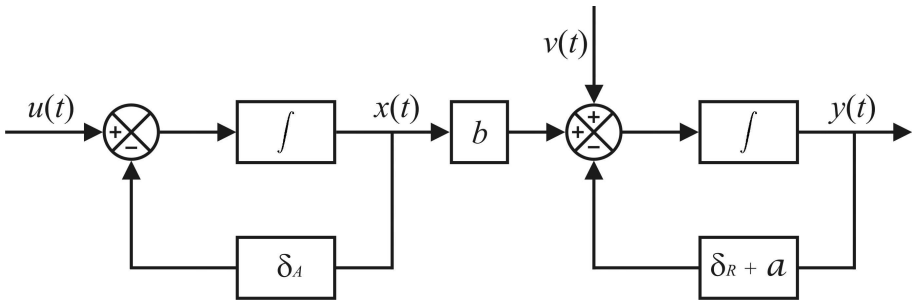


Figura 1.3: *Flujo de señales para el sistema (1.19). $u(t)$ y $v(t)$ son las señales de entrada fluctuantes debidas al ruido intrínseco. La nomenclatura de los distintos componentes de este tipo de representaciones, denominadas* diagramas de bloques*, se detalla en el Apéndice A.*

Efectivamente, la Figura 1.3 muestra el esquema de transmisión de señales que se deriva del sistema de ecuaciones (1.19). A diferencia de lo que sucede con $x(t)$,

que sólo va a tener (en este modelo sencillo) la componente de ruido intrínseco, $y(t)$ ha de responder a una segunda señal de entrada fluctuante: la propia $x(t)$ (a través de la función de transferencia b en el trayecto de entrada). Así que $x(t)$ es precisamente la fuente de ruido extrínseco de $y(t)$. Obsérvese que en el trayecto de realimentación siempre aparece el término de degradación. En el caso de $y(t)$, a la autorregulación inevitable que se produce por el proceso de degradación de la proteína se une la constante a, proveniente de la linealización de la autorregulación transcripcional de R. La suma $\delta_R + a$ recoge así la función análoga que, desde un punto de vista matemático, ejercen una y otra contribución a la realimentación, y a lo cual ya nos hemos referido.

1.3.1. Transmisión de señal

Antes de entrar en detalles sobre la naturaleza de las distintas señales, y como el sistema es lineal[11], vamos a caracterizar las propiedades de propagación que tienen secciones del sistema situadas entre distintos pares de estas señales. Para ello, nos olvidamos por el momento de la naturaleza real de las señales $u(t)$, $v(t)$ y $x(t)$, y las sustituimos por señales-prueba tipo escalón o sinusoidal (Apéndice A). Así, de esta manera, encontramos las propiedades de transmisión que figuran en la Tabla 1.1; como se ve, las que atañen a frecuencias y tiempos característicos sólo dependen del valor de la realimentación A. En particular, obsérvese cómo el tiempo de respuesta t_R del sistema autorregulado disminuye conforme aumenta la fuerza de la autorregulación transcripcional negativa a; o dicho de otro modo, la autorregulación negativa acelera la respuesta del sistema. Esta es una de las propiedades dinámicas básicas que se han propuesto como causa de la selección del control autógeno y que ha podido comprobarse *in vivo* mediante circuitos genéticos sintéticos [53]. En el Capítulo 2 contrastaremos estos resultados con nuestras medidas del tiempo de respuesta de un sistema genético natural autorregulado.

1.3.2. Transmisión de ruido intrínseco

Y ahora entremos en la descripción de las señales estocásticas $u(t)$ y $v(t)$. Como tales señales estocásticas hay que tratarlas en términos estadísticos de promedios y varianzas: por ser fluctuaciones en torno al equilibrio su promedio es nulo, $\langle u(t)\rangle = \langle v(t)\rangle = 0$; y la varianza de estas fluctuaciones las encontraremos por vía de su densidad espectral de potencia, PSD (por *power spectral density*, Apéndice A). Para esto último aprovechemos que estas fluctuaciones son formalmente análogas a las del ruido de disparo (*shot noise*) que surge cuando cargas discretas cruzan las uniones de los semiconductores con tiempos aleatorios [43]. En este sistema electrónico la PSD viene dada por una constante en la frecuencia de valor qI, donde

[11] Es decir, que la respuesta a una suma de señales de entrada es igual a la suma de las respuestas a las señales actuando separadamente (principio de superposición) [43].

entrada → salida	$u(t) \to x(t)$	$v(t) \to y(t)$	$x(t) \to y(t)$
variable TF en la salida	A	R	R
A (autorregulación)	δ_A	$\delta_R + a$	$\delta_R + a$
B	1	1	b
frecuencia característica*	$f_A = \frac{\delta_A}{2\pi}$	$f_R = \frac{\delta_R + a}{2\pi}$	f_R
tiempo de respuesta*	$t_A = \frac{1}{\delta_A}$	$t_R = \frac{1}{\delta_R + a}$	t_R
función de transferencia a $f = 0$	$H_{o,A} = \frac{1}{\delta_A}$	$H_{o,R} = \frac{1}{\delta_R + a}$	$H_{o,R} = b\, H_{o,R}$
función de transferencia	$H_A = \frac{H_{o,A}}{1 + i\frac{f}{f_A}}$	$H_R = \frac{H_{o,R}}{1 + i\frac{f}{f_R}}$	$H = b\, H_R$
ancho de banda equivalente	$\Delta f_A = \frac{\delta_A}{4}$	$\Delta f_R = \frac{\delta_R + a}{4}$	Δf_R

Tabla 1.1: *Propiedades de transmisión para aquellas secciones del sistema situadas entre los pares de señales de entrada/salida que figuran en la primera línea (véase en el Apéndice A la definición de tales propiedades).* A *es la función de transferencia en el trayecto de realimentación;* B*, la función de transferencia en el trayecto de entrada. Nótese la escritura en redondilla de* A *para distinguirla de la concentración de uno de los TFs, A, escrito en cursiva.*
**La frecuencia característica y el tiempo de respuesta se denominan en el Apéndice A de manera genérica como f_o y τ_r, respectivamente.*

q es la unidad discreta del portador de carga e I es la corriente eléctrica promedio y cuyo análogo en nuestro sistema es la AP en el estado estacionario, Π_A para A y $f(A_*, R_*)$ para R según (1.9), que produce en cada caso una unidad discreta de la proteína respectiva [54]. El otro proceso discreto que es fuente de ruido de disparo es la degradación de las proteínas[12]. Como en el estado estacionario este proceso es equivalente al de producción, esto es,

$$\begin{aligned} \Pi_A &= \delta_A A_*, \\ f(A_*, R_*) &= \delta_R R_*, \end{aligned}$$

[12]Fuentes adicionales de ruido intrínseco son los procesos de producción y degradación de los ARN mensajeros. Para tener en cuenta estas contribuciones al ruido, la modelización matemática habría de considerar de forma explícita las ecuaciones de la dinámica del mARN de cada proteína, al modo del sistema (1.7). Un tratamiento de este tipo se desarrolla en [54]. No obstante, nuestro sencillo modelado, basado únicamente en la dinámica de las proteínas, permite llegar de manera cualitativa al mismo tipo de conclusiones sobre el papel de la autorregulación en la transmisión del ruido que se obtendría usando el modelo más complejo. Más adelante comentaremos las modificaciones cuantitativas que, sobre las soluciones proporcionadas por nuestro modelo, introduciría este último.

tenemos las siguientes PSDs asociadas a las respectivas señales $u(t)$ y $v(t)$:

$$\begin{aligned} G_u &= \Pi_A + \delta_A A_* = 2\delta_A A_*, \\ G_v &= f(A_*, R_*) + \delta_R R_* = 2\delta_R R_*. \end{aligned}$$

Al ser constantes para toda frecuencia, las PSDs obtenidas corresponden a sendos ruidos blancos (Apéndice A). Aunque los mecanismos de producción de proteínas restringen el ancho de banda de las PSDs reales, el propio sistema impondrá una limitación aún más severa al contenido frecuencial de las respuestas $x(t)$ e $y(t)$, lo que hace de la asunción de un ruido blanco en estas señales una aproximación apropiada [54].

A partir de la expresión general para la varianza que tenemos en el Apéndice A, dada la forma de las PSDs, y teniendo en cuenta que las varianzas de x e y son las varianzas de la concentraciones en el estacionario ($\sigma_x^2 = \sigma_{A_*}^2$, $\sigma_{y,in}^2 = \sigma_{R_*,in}^2$), obtenemos las siguentes expresiones para la relación ruido-señal $\mathcal{S}$ en cada caso[13],

$$\begin{aligned} \mathcal{S}_{A_*} \equiv \frac{\sigma_{A_*}^2}{A_*} &= 1, \\ \mathcal{S}_{R_*,in} \equiv \frac{\sigma_{R_*,in}^2}{R_*} &= \frac{1}{1+\frac{a}{\delta_R}}. \end{aligned} \tag{1.20}$$

La ecuación (1.20) muestra el efecto de disminución del ruido intrínseco de R_* (respecto al valor de la concentración) conforme aumenta la autorregulación transcripcional a. Por el contrario, esta relación se hace máxima –e igual a la unidad[14], coincidiendo con el caso de A_*– cuando el valor de a se anula, lo cual puede suceder por distintas causas, como veremos enseguida. La Figura 1.5.B es una representación de $\mathcal{S}_{R_*,in}$ en las condiciones que estamos considerando ($\rho_1 > 1$ y $\rho_2 = \rho_3 < 1$). Pues bien, según dicta la expresión (1.14), el valor de a se puede

[13]Obsérvese que hemos añadido el subíndice *in* para señalar que $\sigma_{R_*,in}^2$ es sólo la contribución *intrínseca* al ruido de R_* (debida al ruido de disparo, como hemos visto), para distinguirla de la de carácter extrínseco, $\sigma_{R_*,ex}^2$, que se tratará más adelante. Por tener sólo componente intrínseca, eludimos el uso de subíndice para el ruido de $x(t)$.

[14]Si se tiene en cuenta el ruido de disparo inherente a la producción y degradación de los mensajeros se puede demostrar [54] que las expresiones que rigen el ruido intrínseco son las que siguen:

$$\begin{aligned} \mathcal{S}_{A_*} &= 1 + \mathcal{B}_A, \\ \mathcal{S}_{R_*,in} &= \frac{1+\mathcal{B}_R}{1+\frac{a}{\delta_R}}, \end{aligned}$$

siendo $\mathcal{B}_i$ el parámetro de "ráfaga" (*burst*) del mensajero, $\mathcal{B}_i = \gamma_i/\delta_{m_i}$ ($i = A, R$), que es una medida del promedio de proteínas producidas por cada transcrito [55, 56]. Como se ve, estas expresiones introducen en (1.20) la corrección del factor $(1+\mathcal{B}_i)$. La corrección implica además que sea este factor el que dé el valor máximo que alcanza el ruido en circuitos carentes de autorregulación negativa.

minimizar haciendo que el factor de transcripción R no tenga afinidad alguna por el sitio operador ($k_R \to \infty$), lo que anula las probabilidades P_{S2} y P_{S3}; pero también se minimiza si, en el extremo contrario, se consideran afinidades muy grandes para el represor ($k_R \to 0$), lo que hace que $R_*/R_o \to \rho_2 = \rho_3$, anulando así el valor de los paréntesis de la expresión (1.14). En este último caso, el operador está saturado de represor y la AP se hace insensible a pequeños desplazamientos en las concentraciones. Como se ve en la Figura 1.5.B, para cada k_A dado el ruido intrínseco se hace mínimo (con el consiguiente máximo de autorregulación) en la zona de valores intermedios de k_R (región de máxima sensibilidad a los cambios en la concentración del represor).

1.3.3. Transmisión de ruido extrínseco

Hemos visto que el ruido extrínseco de R_* es debido a la transmisión de las fluctuaciones del estado estacionario de A sobre el circuito regulado[15]. La PSD de la señal de fluctuación externa, $G_x(f)$, ya no es la de un ruido blanco por ser $x(t)$, a su vez, una señal ya filtrada por el circuito de A[16]:

$$G_x(f) = |H_A(f)|^2 G_u.$$

Obsérvese en la Figura 1.3 y en la Tabla 1.1 cómo la función de transferencia en el trayecto de entrada (B en la nomenclatura del Apéndice A) es igual a b, con lo que la PSD del ruido extrínseco de R_* viene dada por

$$G_{y,ex}(f) = |H_R(f)|^2 b^2 G_x(f) = |H_R(f)|^2 b^2 |H_A(f)|^2 G_u,$$

resultado equivalente a la propagación de la PSD del ruido blanco $u(t)$ a través del circuito completo, cuya función de transferencia combinada sería, por tanto, $H_c(f) \equiv H_A(f)\, b\, H_R(f)$. A partir de su PSD calculamos la varianza de este ruido extrínseco (Apéndice A)[17],

$$\sigma^2_{R_*,ex} = \int_{-\infty}^{\infty} G_{y,ex}(f)\, \mathrm{d}f = b^2 |H_{o,A}|^2 |H_{o,R}|^2 G_u\, 2 \int_0^{\infty} \frac{1}{1+\left(\frac{f}{f_A}\right)^2} \frac{1}{1+\left(\frac{f}{f_R}\right)^2}\, \mathrm{d}f,$$

[15] En realidad la existencia de regulación transcripcional de A sobre R tiene influencia en el propio ruido intrínseco $\sigma^2_{R_*,in}$. No obstante, esto sucede esencialmente a través de la concentración en el estacionario de A y no tanto por las fluctuaciones $x(t)$. Recordemos que la PSD de este ruido intrínseco es proporcional a la AP de R en el estacionario, esto es, a $f(A_*, R_*)$. Las fluctuaciones en torno al equilibrio de A y R implicarían tan sólo correcciones de segundo orden en el cálculo de la PSD.

[16] Sobre la transmisión cuadrática de las PSDs, consúltese el Apéndice A.

[17] Nótese de nuevo que la varianza de la perturbación es lo mismo que el ruido del estacionario, $\sigma_{R_*,ex} = \sigma_{y,ex}$.

donde hemos usado que la función es par en f. Sustituyendo los valores de G_u y las funciones de transferencia a frecuencia cero (Tabla 1.1), obtenemos, finalmente, la siguiente expresión para el ruido extrínseco de R_*:

$$\sigma^2_{R_*,ex} = C\,\frac{b^2}{(\delta_R + a)^2}\,A_* = C\,|H_o|^2\,\sigma_x^2, \tag{1.21}$$

donde se ha tenido en cuenta que $\sigma_x^2 = A_*$. C es la constante dada por

$$C = \frac{4}{\delta_A}\int_0^\infty \frac{1}{1+\left(\frac{f}{f_A}\right)^2}\frac{1}{1+\left(\frac{f}{f_R}\right)^2}\,\mathrm{d}f. \tag{1.22}$$

Si la autorregulación negativa es muy grande ($\delta_R + a \gg \delta_A$), entonces $f_R \gg f_A$ y el valor de C se aproxima a la unidad, pues en este caso

$$C_{a\gg 0} \sim \frac{4}{\delta_A}\int_0^\infty \frac{1}{1+\left(\frac{f}{f_A}\right)^2}\,\mathrm{d}f = 1,$$

con lo que el valor del ruido extrínseco quedaría como sigue:

$$\sigma^2_{R_*,ex}|_{a\gg 0} = \frac{b^2}{(\delta_R + a)^2}\,A_* = |H_o|^2\,\sigma_x^2. \tag{1.23}$$

Este último resultado es fácilmente interpretable: la señal $x(t)$ tiene un espectro de frecuencias con un ancho de banda Δf_A. Al ser más permisivo el segundo filtro por su mayor ancho de banda, va a dejar pasar todas las frecuencias del ruido de $x(t)$, que son reescaladas por el cuadrado de la altura del segundo filtro a frecuencias bajas, y por ello, en términos absolutos, obtenemos el mismo reescalamiento para el ruido σ_x^2 en su totalidad.

De todas formas, la constante C se puede calcular para el caso general[18] y vale:

$$C = \frac{1}{1+\frac{\delta_A}{\delta_R+a}},$$

con lo que la expresión del ruido extrínseco de R_* queda como sigue:

$$\sigma^2_{R_*,ex} = \frac{b^2}{(\delta_R + a)(\delta_R + a + \delta_A)}\,\sigma_x^2, \tag{1.24}$$

lo que implica que también el ruido extrínseco es mayor cuanto menor es la autorregulación transcripcional a. Aunque, como vimos más arriba, el valor de a

[18]La integral (1.22) se evalúa por descomposición en fracciones simples con el cambio de variable $s = f^2$. Por ejemplo, véase [57] (integral tipo E-2).

puede minimizarse ante circunstancias extremas de regulación muy fuerte o muy débil por parte de R, la obtención de un ruido extrínseco máximo va a requerir, en las condiciones consideradas[19], la anulación de la probabilidad de enlace del represor ($k_R \to \infty \Rightarrow P_{S2} = P_{S3} = 0$), lo cual, a la vez que minimiza a, maximiza la expresión de b (1.13). En este caso, la fórmula del ruido extrínseco (1.24) se reduce a la siguiente:

$$\sigma^2_{R_*,ex}\big|_{k_R\to\infty} = \frac{b^2}{\delta_R(\delta_R+\delta_A)}\,\sigma^2_x.$$

No obstante, la eliminación de la autorregulación negativa ($a = 0$) por sí sola no es suficiente para tener una propagación máxima del ruido de $x(t)$, puesto que esta propagación depende aún, a través del parámetro de entrada b, de la concentración estacionaria del activador y de la afinidad de éste por la zona promotora. En ausencia de regulación por R, la expresión de b en función de las probabilidades de los estados se reduce ahora a la siguiente:

$$b\big|_{k_R\to\infty} = m\,\delta_R\,\frac{R_*}{A_*}\left(\frac{\rho_1}{R_*/R_o}-1\right)P_{S1} = m\,\delta_R\,\frac{R_o}{A_*}\,(\rho_1-1)\,\frac{\theta}{(1+\theta)^2}, \qquad (1.25)$$

donde la variable $\theta \equiv \left(\frac{A_*}{k_A}\right)^m$ recoge la doble dependencia mencionada. ¿Cuál es el valor de θ que maximiza el ruido extrínseco? Fijémonos en que la expresión (1.25) es siempre positiva si A es activador ($\rho_1 > 1$) y negativa si es represor ($\rho_1 < 1$). Por tanto, el valor de θ que maximiza la función $g(\theta) \equiv \theta/(1+\theta)^2$ también maximiza $\sigma^2_{R_*,ex}$. Este valor extremo se obtiene para $\theta = 1$, esto es, para $A_* = k_A$ (Figura 1.4.A) y vale

$$\text{máx}\left(\sigma^2_{R_*,ex}\right) = \frac{1}{1+\delta_A/\delta_R}\frac{1}{A_*}\left[\frac{mR_o}{4}\,(\rho_1-1)\right]^2.$$

Obsérvese que si A es represor, el valor máximo de ruido que se puede alcanzar está acotado y se obtiene para $\rho_1 = 0$. En cambio, si A es activador se pueden alcanzar valores de ruido mucho mayores, tan sólo limitados por la magnitud que pueda alcanzar ρ_1 en los sistemas celulares. El ruido extrínseco se anula en $\rho_1 = 1$ puesto que entonces la unión de A no afecta en nada a la AP, lo que implica la anulación del parámetro de entrada b en (1.25)[20].

En (1.21) y (1.25) hemos visto que encontrar el máximo del ruido extrínseco equivale a encontrar el máximo del valor absoluto de b. Como b es la derivada

[19] Recordemos: $\rho_1 > 1$, $\rho_2 = \rho_3 < 1$. La Figura 1.5.C es una representación del ruido extrínseco (normalizado por R_*) con valores para las ρ's que satisfacen estas condiciones.

[20] Con el represor incapaz de unirse al promotor ($k_R \to \infty$) y con $\rho_1 = 1$, la concentración estacionaria de R sería entonces $R_* = R_o$ –véase, más abajo, la expresión (1.27) para el estacionario R_* en ausencia de regulación por el represor.

respecto a A de la AP, que en el caso sin regulación debida a R se reduce a

$$f(A_*) = \Pi_R \frac{1+\rho_1\theta}{1+\theta},$$

el valor de $\theta = 1$, que maximiza el valor absoluto de la pendiente b, marca la concentración de A a la que la AP es más sensible a los cambios en esta concentración y por ello el valor de la concentración del regulador que iguala a su tasa de disociación es el punto crítico en el que la concentración de R va a ser más sensible al ruido de A.

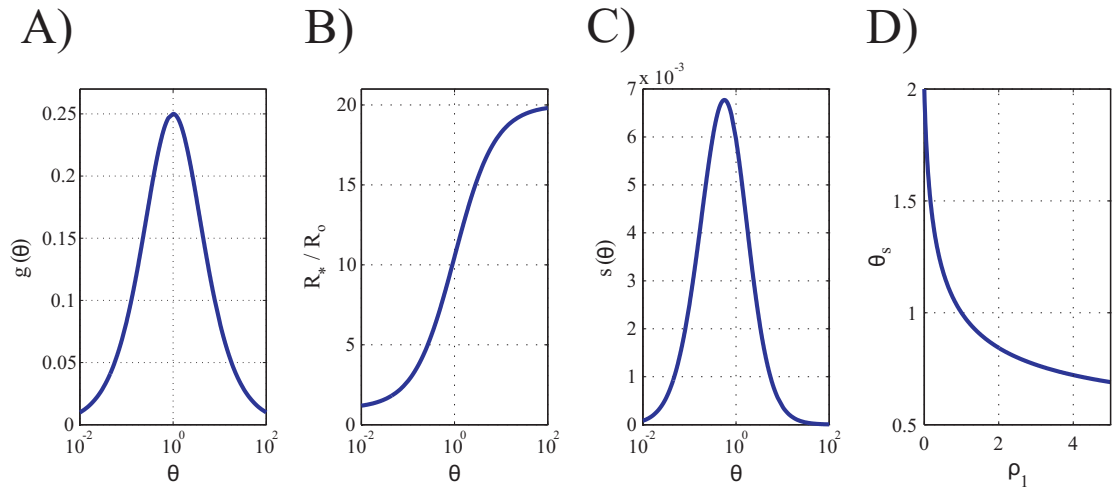


Figura 1.4: *Estas cuatro gráficas están calculadas en ausencia de regulación por R. A) El ruido es proporcional al cuadrado la función $g(\theta)$ aquí representada y tiene su máximo en $\theta = 1$. B) Estado estacionario de R en función de θ para $\rho_1 = 20$. C) Función $s(\theta)$ (proporcional a la relación ruido-señal extrínseco en ausencia de regulación por R, véase el texto) para $\rho_1 = 20$. Su máximo se desplaza a la izquierda de $\theta = 1$ por ser $\rho_1 > 1$. D) Valores de θ que maximizan $s(\theta)$ en función de ρ_1 (véase la nota 21).*

Este máximo de ruido lo es en términos absolutos, es decir, sin normalizarlo por el valor de la señal a la que afecta. En cambio para la relación ruido-señal en ausencia de autorregulación,

$$\mathcal{S}_{R_*,ex}\big|_{k_R\to\infty} = \frac{\sigma^2_{R_*,ex}}{R_*}\Big|_{k_R\to\infty} = \frac{A_*}{R_*}\frac{b^2}{\delta_R(\delta_R+\delta_A)}, \tag{1.26}$$

el máximo no coincide en general con $A_* = k_A$. Ello es debido a que la concentración en el estacionario del represor también depende de θ,

$$\frac{R_*}{R_o} = \frac{f(A_*)}{\Pi_R} = \frac{1+\rho_1\theta}{1+\theta}, \tag{1.27}$$

que es una función decreciente para $\rho_1 < 1$ y creciente en θ si $\rho_1 > 1$ (Figura 1.4.B). Escrita en función de θ, la expresión (1.26) queda como sigue:

$$\mathcal{S}_{R_*,ex}\big|_{k_R\to\infty} = K\frac{g^2}{R_*/R_o} = K\frac{\theta^2}{(1+\rho_1\theta)(1+\theta)^3}, \quad K \equiv \frac{1}{1+\delta_A/\delta_R}\frac{R_o}{A_*}\, m^2\,(\rho_1-1)^2\,.$$

El valor θ_s que maximiza la función $s(\theta) \equiv \frac{\theta^2}{(1+\rho_1\theta)(1+\theta)^3}$ (Figura 1.4.C) es menor que la unidad si $\rho_1 > 1$ (con un mínimo asintótico en $1/2$ para valores muy grandes de ρ_1), y mayor que la unidad si $\rho_1 < 1$ (con un valor máximo de 2 para $\rho_1 = 0$) tal y como se ve en la Figura 1.4.D[21]. Este efecto también puede observarse en la Figura 1.5.C, con $\rho_1 = 10$ y $\rho_2 = \rho_3 = 0.2$, donde el pico máximo de $\mathcal{S}_{R_*,ex}$ está desplazado a la izquierda de la unidad.

1.3.4. Ruido total

Como el sistema es lineal, para obtener el ruido total de R_* no hay más que sumar las contribuciones intrínseca y extrínseca:

$$\mathcal{S}_{R_*,tot} = \frac{\sigma^2_{R_*}}{R_*} = \frac{1}{1+\frac{a}{\delta_R}} + \frac{1}{G}\frac{b^2}{(\delta_R+a)(\delta_R+a+\delta_A)},$$

siendo G la ganancia entre las concentraciones en el estacionario $G = R_*/A_*$.

La Figura 1.5.D muestra que, con los valores de los parámetros $\rho' s$ utilizados en ella, el comportamiento global del ruido está fuertemente determinado por el del ruido extrínseco (Figura 1.5.C). Finalmente, mencionemos que la disminución del ruido de las concentraciones asociada a la autorregulación negativa se ha comprobado experimentalmente mediante la construcción de pequeños circuitos genéticos sintéticos [58].

[21] La curva trazada en la Figura 1.4.D es descrita por la función continua siguiente:

$$\theta_s = \begin{cases} 2 & \text{si } \rho_1 = 0 \\ \frac{\rho_1-1}{4\rho_1}\left(1-\sqrt{1+16\frac{\rho_1}{(\rho_1-1)^2}}\right) & \text{si } 0<\rho_1<1 \\ 1 & \text{si } \rho_1 = 1 \\ \frac{\rho_1-1}{4\rho_1}\left(1+\sqrt{1+16\frac{\rho_1}{(\rho_1-1)^2}}\right) & \text{si } \rho_1 > 1 \end{cases}$$

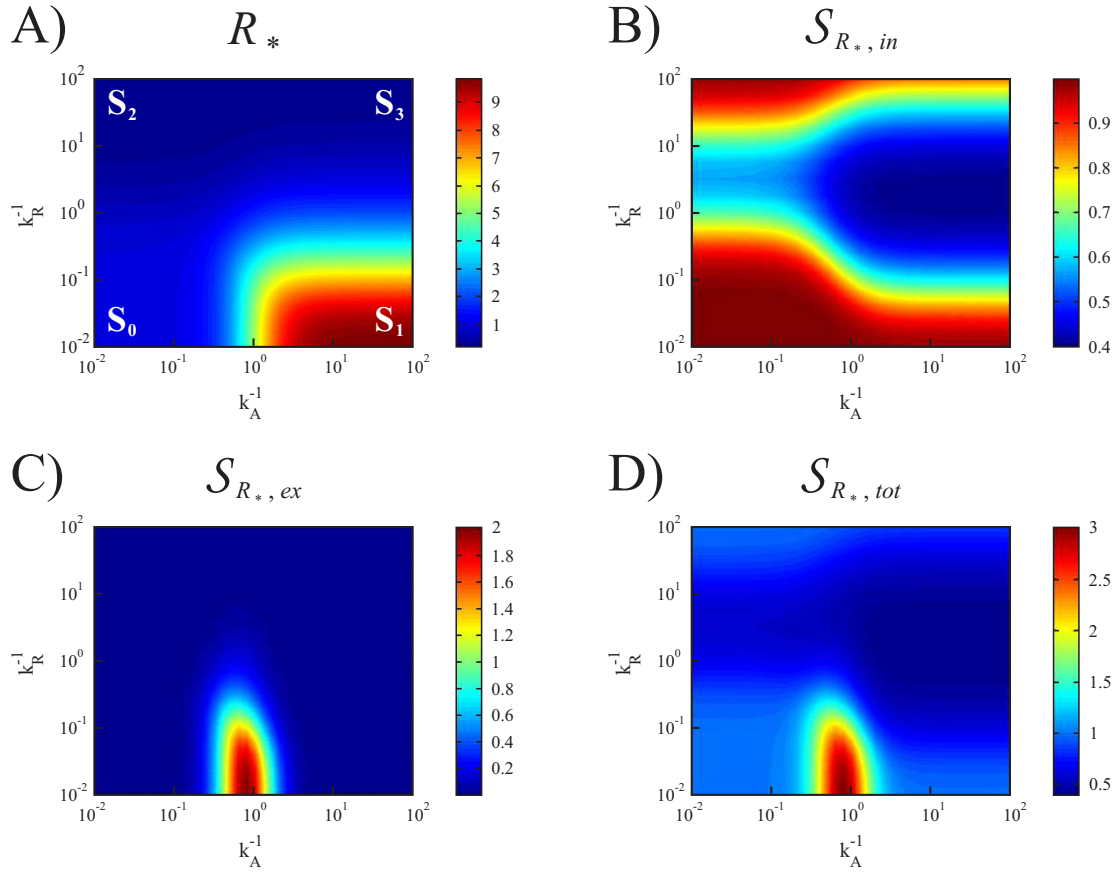


Figura 1.5: *Representación de la concentración del represor y las distintas componentes del ruido de su concentración frente a las inversas de k_A y k_R –cuanto mayor es la inversa de una constante de disociación, más fuerte es la regulación por parte del factor de transcripción correspondiente. A) Concentración estacionaria del represor, R_*. Las cuatro esquinas corresponden a situaciones extremas de regulaciones muy fuertes o muy laxas. En cada caso, está anotado el estado del promotor que domina. B) Ruido intrínseco de R_*: para cualquier valor dado de la afinidad del activador el mínimo de $\mathcal{S}_{R_*,in}$ se localiza en valores intermedios de k_R. C) Ruido extrínseco de R_*: el máximo está situado en la zona en la que la afinidad de R es muy baja y la de A es intermedia. D) Ruido total de R_*. Parámetros: $A_* = R_o = 1$, $\delta_A = \delta_R$, $m = n = 2$, $\rho_1 = 10$, $\rho_2 = \rho_3 = 0.2$.*

Capítulo 2

La autorregulación en la optimización de la respuesta de un sistema genético natural

En este capítulo se presentan los resultados del siguiente artículo, incluido en el Apéndice B:
Camas FM, Blázquez J, Poyatos JF (2006) Autogenous and nonautogenous control of response in a genetic network. Proc Natl Sci USA 103: 12718-12723

En el Capítulo 1 hemos presentado un bosquejo de las principales propiedades de la autorregulación negativa, las cuales fueron obtenidas a partir de un circuito transcripcional mínimo en el que había poco más que el propio fenómeno de la autorregulación. En este simple escenario la autorregulación negativa como sistema de control presenta una serie de ventajas (rapidez de respuesta y menor ruido) frente a los sistemas no autorregulados. Aunque estas propiedades habían sido confirmadas en experimentos que involucraban a circuitos mínimos sintéticos [53, 58], restaba por ver en qué medida las redes naturales se encontrarían precisamente en un régimen en el que estas ventajas se pusieran de manifiesto. La cuestión no era baladí, porque en los sistemas vivos la autorregulación está embebida en un entorno de interacciones complejo, que ni siquiera se limita a la maraña de la red global de regulación transcripcional. Para estudiar los beneficios del control autógeno sobre un sistema genético natural escogimos el sistema SOS de respuesta inducible frente a daños en el ADN de *Escherichia coli* [59], cuyo regulador maestro, LexA, autorregula su propia transcripción. En este estudio hemos seguido una estrategia que combina experimentación y modelización matemática; en el plano experimental se ha rediseñado el circuito natural de *E. coli* para tener una variante del mismo en la que LexA no se autorregulara y así poder contrastar la dinámica *in vivo* del SOS en presencia y ausencia de autorregulación.

2.1. El sistema SOS

2.1.1. ¿Por qué el SOS?

Aparte de por la presencia del control autógeno, la elección del sistema SOS obedece a varias razones adicionales: i) el SOS es un módulo funcional, lo que quiere decir que involucra a un conjunto bien delimitado de especies de moléculas de cuyas interacciones mutuas emerge una función separable del resto de funcionalidades celulares [5]; ii) existe una descripción cualitativa de esta funcionalidad bastante completa, lo que da pie al estudio cuantitativo de la misma a través de la modelización matemática [60]; iii) en cuanto a la topología de la subred de transcripción involucrada, veremos que el SOS ejemplifica uno de los varios tipos de motivos que se dan con recurrencia en la red de transcripción [9, 10]; y iv) el funcionamiento del SOS no está determinado únicamente por estas relaciones transcripcionales [59], lo cual complica la dinámica del sistema.

Entremos ahora en detalles sobre todo esto, empezando por una breve descripción del funcionamiento del SOS.

2.1.2. El sistema SOS: módulo funcional

El sistema SOS consta de un represor maestro que, aparte de autorregular su propia transcripción, inhibe la de una batería de proteínas relacionadas con la respuesta al daño en el ADN [59]. En particular, una de estas proteínas, RecA, constituye junto con LexA el módulo genético de control fundamental de la respuesta SOS. Diferentes agentes externos como la radiación ultravioleta (UV) son capaces de generar daño en el ADN, dejando expuestos fragmentos de hebra simple. La unión a estos fragmentos de la proteína RecA conforma el complejo denominado RecA*. La presencia de RecA* favorece la autoproteólisis (degradación) de LexA, lo que levanta la represión de los genes SOS encargados de la reparación del ADN. Una vez que el ADN es reparado, el nivel de RecA* decae y con él la autodigestión de LexA, lo cual en última instancia restituye el estado de no-inducción del sistema.

2.1.3. El sistema SOS: motivo de red

Los motivos de red son patrones de conexión o pequeños subgrafos que aparecen de una manera recurrente dentro de una red. La definición de motivo exige que la frecuencia de aparición de los mismos sea estadísticamente significativa en relación a un modelo nulo en el que las conexiones entre los nodos de la red son recableadas de una manera aleatoria [9, 61, 62]. En el ámbito de la Biología, los motivos fueron descritos por primera vez en la red transcripcional de *E. coli* por Uri Alon y sus colaboradores [9]. De hecho, a pesar de que también han sido con-

sideradas otras redes como las neuronales y ecológicas [61] o las de interacciones entre proteínas [63], el estudio de los motivos se ha ceñido sobre todo al ámbito de la regulación transcripcional.

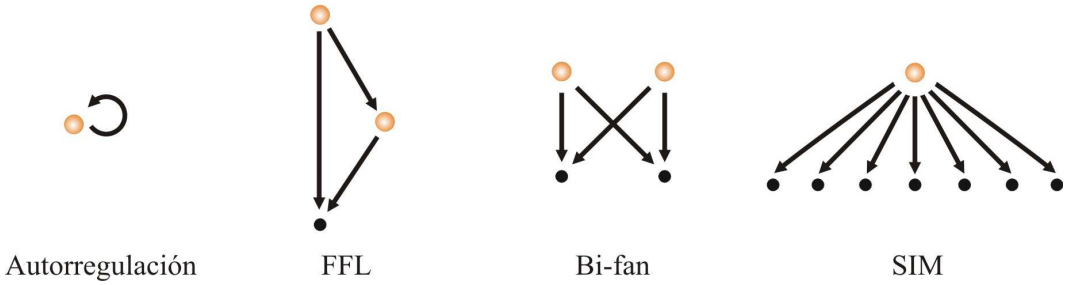


Figura 2.1: *Motivos de red transcripcionales. Los nodos coloreados representan a los genes que codifican factores de transcripción.*

Y aunque dependiente de las hipótesis nulas subyacentes al modelo aleatorio [64, 65], la tipología de los motivos que se han venido identificando en *E. coli* y otras bacterias, en levaduras [66] e incluso en organismos superiores [67, 68] no suele apartarse de la que aparece en la Figura 2.1 (lo cual resulta, cuanto menos, llamativo). Como se ve, a la propia autorregulación, que es motivo de red [42, 69], le acompañan el *feed-forward loop*[1] (FFL), el bi-fan y el módulo de entrada simple (SIM, por *Single Input Module*).

A la arquitectura de cada uno de estos motivos se le atribuye tareas específicas de procesamiento de información [10]; bajo esta visión, los motivos son unidades funcionales autónomas de carácter adaptativo. En el próximo capítulo veremos que este escenario evolutivo dista mucho de recibir un respaldo unánime, existiendo una corriente alternativa de opinión que aboga por la selección neutral de los motivos. Las críticas al carácter determinante que para la funcionalidad tendría la estructura del motivo se centran principalmente en dos aspectos: en primer lugar, ya sólo ateniéndonos a las conexiones transcripcionales, el comportamiento de los motivos podría depender en gran medida del contexto que supone la red en la que están embebidos [70] y no tanto de la estructura local constituida por el propio motivo [13, 14, 16]. En segundo lugar, puesto que los subgrafos transcripcionales sólo capturan un aspecto de la maquinaria reguladora, la funcionalidad inferida a partir de ellos puede verse fuertemente matizada por la existencia de interacciones postranscripcionales y postraduccionales [13].

En el Capítulo 3 estudiaremos los determinantes que pueden llevar al ensamblaje –neutral o selectivo– de todas estas arquitecturas. De momento, prestemos una especial atención al motivo SIM, caracterizado por una batería de operones que están bajo el control de un mismo factor de transcripción o regulador maestro

[1] Mantenemos la nomenclatura anglosajona.

(Figura 1). Sólo cuando este patrón de conexiones implica a un número grande de regulaciones la señal estadística es significativa y podemos hablar con propiedad de motivo SIM [9] (véase el Capítulo 3). Además, en rigor, la definición de SIM exige que dicho control tenga el mismo carácter –de activación o de represión– para todos los operones regulados y que sobre éstos no exista más regulación transcripcional que la que les viene del regulador maestro [9].

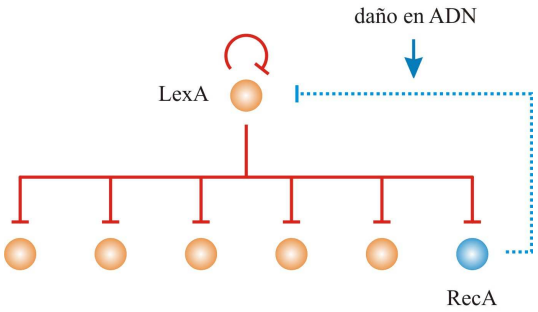


Figura 2.2: *Esquema de conexiones del sistema SOS.*

La relación de todo esto con el sistema SOS estriba en que, en lo que atañe a sus conexiones transcripcionales, el SOS dibuja un arquetípico motivo SIM con el regulador maestro autorregulado, circunstancia ésta que se da con frecuencia en esta clase de motivo [10]. Efectivamente, la Figura 2.2 abstrae los detalles del funcionamiento del SOS reduciéndolo a un esquema de conexiones de distinta naturaleza: por un lado, tenemos las conexiones de represión transcripcional –en rojo y que incluyen la autorregulación de LexA–; y por otro, la inhibición de tipo postraduccional que hemos esquematizado con la conexión punteada azul, conexión que sólo se "activa" en presencia de daños en el ADN. La modelización matemática basada en este sencillo esquema será suficiente para parametrizar e interpretar nuestros resultados experimentales.

Como veremos en el Capítulo 3, la subred del SOS exhibe un alto grado de aislamiento del resto de la red en general, lo cual contribuye a que, en este caso, se pueda asimilar el motivo a una entidad funcional autónoma. No obstante, al hilo de la discusión anterior, este es un claro ejemplo de cómo la estructura transcripcional no basta para dar cuenta de la funcionalidad del sistema. Efectivamente, la presencia de la conexión postraduccional implica que sólo en ausencia de daño la dinámica del SOS puede ser inferida a partir del esquema de regulación transcripcional SIM. De todas formas, comenzaremos con la modelización matemática del sistema en estas circunstancias, puesto que el análisis de su solución estacionaria constituye un buen punto de partida tanto para la comprensión del sistema en situaciones dinámicas más complejas como para la planificación del rediseño experimental del control autógeno del regulador maestro, LexA. Más adelante nos ocuparemos de la modelización del sistema en presencia de daño.

2.1.4. Dinámica del sistema SOS en ausencia de daño

En este caso la modelización de la dinámica es tanto o más sencilla que la de los circuitos del Capítulo 1. Así, siendo x e y las concentraciones respectivas de LexA y alguna de las otras proteínas reguladas por LexA (digamos, por ejemplo, RecA), tendríamos las siguientes ecuaciones:

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \frac{\alpha_x}{1+\frac{x}{k_x}} - \delta_x x, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= \frac{\alpha_y}{1+\frac{x}{k_y}} - \delta_y y, \end{aligned} \tag{2.1}$$

donde los parámetros tipo δ, k y α dan cuenta respectivamente de las tasas de degradación, las constantes de disociación del represor y la actividad del promotor (AP) [71] en ausencia de represor[2]. La solución de este sistema es, por supuesto, el conjunto de concentraciones estacionarias (x_*, y_*) en el estado sin inducir:

$$\begin{aligned} x_* &= \frac{\alpha_x}{\delta_x}\frac{1}{1+\frac{x_*}{k_x}} = \frac{k_x}{2}\left(\sqrt{1+\frac{4\alpha_x}{k_x\delta_x}}-1\right), \\ y_* &= \frac{\alpha_y}{\delta_y}\frac{1}{1+\frac{x_*}{k_y}}. \end{aligned}$$

El sistema (2.1) plantea una situación similar a la tratada en el Capítulo 1 y, por tanto, como sucedía entonces, si se analizara el papel que juega la autorregulación en el estado estacionario introduciendo fuentes de ruido volveríamos a encontrar que la autorregulación disminuye la varianza de las fluctuaciones en la concentración de LexA.

La presencia de la autorregulación condiciona a priori el nivel absoluto de la concentración de LexA en el estado estacionario, puesto que, si no existiese sitio alguno de unión al ADN (*binding site*, BS) para este regulador en la zona promotora que antecede al gen que lo codifica, tal concentración sería mayor (en concreto, valdría $x_* = \alpha_x/\delta_x$). Pero, partiendo de que el nivel estacionario debe estar sujeto a un proceso de selección adaptativa, la combinación de un promotor fuerte con la autorregulación negativa hay que evaluarla como un todo, y por ello, la alternativa al promotor autorregulado no es este mismo promotor sin el BS para LexA, sino otros promotores en los que una autorregulación más débil se compensara con un valor de α también más pequeño para así dar lugar a una misma concentración estacionaria en el estado sin inducir. Si consideramos el caso extremo de ausencia de toda autorregulación ($k_x \to \infty$), la transcripción se produce entonces con una

[2]La AP de x es en general toda la expresión $\Pi(x) \equiv \alpha_x/(1+\frac{x}{k_x})$. En ausencia de represor tenemos la AP "en vacío", $\Pi(0) = \alpha_x$, que es la cota máxima de $\Pi(x)$. Y lo mismo vale para la AP de y.

tasa α_{na} constitutiva[3] y la ecuación para la concentración de LexA es

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \alpha_{na} - \delta_x x \quad \Rightarrow \quad x_{*,na} = \alpha_{na}/\delta_x,$$

donde se ve que se obtendría el mismo valor del estado estacionario del circuito original ($x_{*,na} = x_{*,au}$) si

$$\frac{\alpha_{na}}{\alpha_{au}} = \left(1 + \frac{x_{*,au}}{k_x}\right)^{-1}, \tag{2.2}$$

lo cual implica necesariamente que el promotor constitutivo ha de ser más débil que el autorregulado ($\alpha_{na} < \alpha_{au}$).

2.2. Rediseño del control autógeno

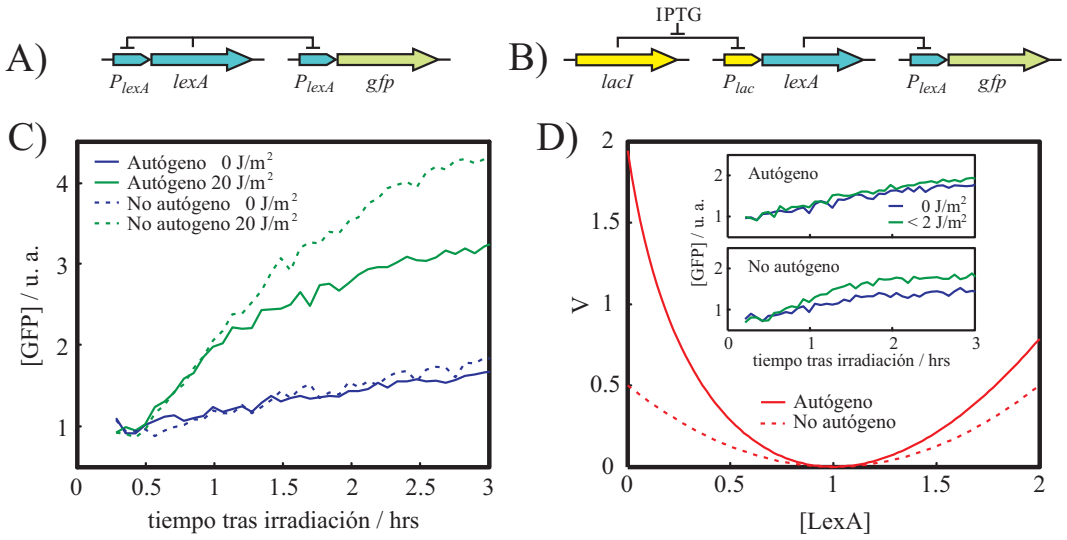


Figura 2.3: *Control autógeno y no autógeno de la respuesta SOS. A) Circuito LexA natural autorregulado más gen reportero gfp. B) Circuito sintético LexA no autorregulado más gen reportero gfp. C) Dinámica SOS (fluorescencia GFP) siguiendo a la irradiación UV de células portadoras de uno u otro circuito. Niveles GFP sin radiación en azul. D) Potenciales de estabilidad asociados al estado estacionario en ausencia de daño para ambos circuitos –véanse las ecuaciones (2.4). Concentraciones de LexA en unidades del estado sin inducir de manera que $x_* = 1$. Encarte: Dinámica SOS para dosis UV muy bajas (< 2 J/m^2). Se manifiesta el incremento de estabilidad asociado al control autógeno. La concentración de GFP siempre en unidades arbitrarias.*

[3]Subíndice *na* por *no autoregulado*. Por simetría, a partir de ahora denotaremos por α_{au} a la tasa α_x del caso autorregulado.

Efectivamente, el rediseño experimental del control de LexA se hizo bajo el presupuesto de mantener en el circuito carente de autorregulación la concentración de este represor en el estado sin inducir. Por tanto, la construcción del sistema sintético no podía limitarse a la eliminación del sitio de enlace de LexA, puesto que, como acabamos de ver, ello dispararía el valor de dicha concentración.

En lugar de esto planteamos la confrontación de los dos circuitos que aparecen en la Figura 2.3. Frente a la autorregulación exhibida por el gen *lexA* en el circuito natural (Figura 2.3.A), en el rediseño sintético *lexA* se halla bajo el control de un promotor reprimible por LacI (Figura 2.3.B). A su vez, la represión ejercida por LacI, y, por tanto, la producción de LexA, puede regularse de forma externa mediante la molécula inductora isopropil-β-D-tiogalactopiranósido (IPTG). De este modo, diferentes concentraciones de IPTG producen diferentes tasas de transcripción en la estirpe sintética, lo cual, de cara al estudio dinámico que pretendíamos, era equivalente a la modulación de la fuerza del promotor no autorregulado, es decir, a la modulación del valor de la constante α_{na}.

Para poder medir las concentraciones de LexA que se producen en las estirpes portadoras de uno u otro circuito se introdujo en ambas un mismo plásmido en el que el gen codificante de la proteína de fluorescencia verde GFP (por *green fluorescent protein*) está bajo el control de la zona operadora de *lexA*. Debido a esto último, la dinámica de la concentración z de la GFP se puede aproximar por

$$\frac{\mathrm{d}z}{\mathrm{d}t} = \frac{\alpha_{au}}{1 + \frac{x}{k_x}} - \delta_z z. \tag{2.3}$$

Nótese que hemos aproximado la AP "en vacío" de la GFP por la de LexA ($\alpha_z = \alpha_{au}$), basándonos en el hecho de que los genes *lexA* y *gfp* están bajo el control de un mismo promotor. Aunque en la realidad existirán diferencias en el número de proteínas que son liberadas en uno y otro caso, las tasas de producción van a ser cuanto menos proporcionales. Y con esto nos basta a la hora de cuantificar el sistema, puesto que la concentración de proteína GFP se expresa en unidades arbitrarias. Por otro lado, la equiparación de las constantes de disociación de LexA ($k_z = k_x$) está aún más justificada, pues los sitios de unión reconocidos por LexA y su entorno en el ADN tienen exactamente la misma secuencia de nucleótidos en ambos casos.

La ecuación (2.3) para la dinámica de la GFP por un lado, y las medidas en el fluorímetro por otro, son el punto de partida para la determinación de la dinámica y los parámetros del sistema [71]. La manera rigurosa de hacerlo se encuentra explicada en detalle en la sección de *Materiales y métodos* del artículo del Apéndice B. No obstante, de manera sucinta diremos que la evolución de la concentración z se puede conocer con una buena resolución temporal (en intervalos temporales de unos 4 minutos) a través de medidas duales de fluorescencia y crecimiento (densidad óptica, DO) de los cultivos bacterianos correspondientes. A partir de la derivada de la concentración z podemos calcular la actividad del

promotor $\Pi(z)$ para la GFP, que será tanto menor cuanto más represor LexA exista en el medio. Y así, de esta manera, se puede reconstruir en última instancia la dinámica de LexA bajo uno u otro sistema de control.

2.3. Resultados experimentales

2.3.1. Sobreexpresión y estabilidad

Una vez implementados los circuitos, en primer lugar estudiamos la respuesta de ambos sistemas a una misma dosis de radiación UV. Nuestras medidas muestran cómo el sistema carente de control autógeno exhibe una respuesta más acusada (Figura 2.3.C). La disminución de la concentración de LexA en presencia de daño en el ADN es compensada en el circuito autorregulado con un aumento de la producción de este factor de transcripción. Dicha compensación no existe en el circuito sin autorregulación, en el cual la tasa de transcripción es indiferente a los cambios en los niveles de LexA. Por tanto, en condiciones de daño, el equilibrio entre producción y degradación se alcanza en el circuito autorregulado a un nivel relativamente más alto de LexA, lo cual se traduce en una mayor represión de la producción de GFP en la estirpe portadora de este circuito.

La autorregulación impide que la concentración de represor en el estado inducido se aleje demasiado de la correspondiente al estado sin inducir, y por tanto, cuando tal autorregulación está presente, provee de una ganancia en estabilidad a este estado de partida[4]. Un modo alternativo y más visual de entender este incremento de la estabilidad es mediante la representación de la dinámica en torno al estado estacionario en términos de pozos de potencial [72, 73]. Bajo este formalismo, el sistema de primer orden para la concentración de LexA, $\mathrm{d}x/\mathrm{d}t = F(x)$, se considera análogo al de una partícula fuertemente amortiguada[5] situada en el fondo de un pozo de potencial $V(x)$ tal que $F(x) = -\mathrm{d}V(x)/\mathrm{d}x$. Integrando las dinámicas del caso autorregulado y sin autorregular obtenemos las expresiones de los pozos de potencial respectivos:

$$\begin{aligned} V(x)_{na} &= \frac{\delta_x}{2}(x^2 - x_*^2) - \alpha_{na}(x - x_*) = \frac{\delta_x}{2}(x - x_*)^2, \\ V(x)_{au} &= \frac{\delta_x}{2}(x^2 - x_*^2) - \alpha_{au}k\left[\log\left(1 + \frac{x}{k_x}\right) - \log\left(1 + \frac{x_*}{k_x}\right)\right], \end{aligned} \quad (2.4)$$

donde se han ajustado las constantes de los potenciales de manera que $V(x_*)_{na} = V(x_*)_{au} = 0$ (Figura 2.3.D).

[4]Lo cual es otra manera de leer los resultados de reducción del ruido o fluctuaciones de la concentración que fueron tratados en el Capítulo 1.

[5]Con la fuerza proporcional a la velocidad y no a la aceleración.

La estabilidad del estado estacionario viene dada por el valor de la segunda derivada del potencial evaluada en el dicho estado, $\sigma = \mathrm{d}F(x)/\mathrm{d}x|_{x=x_*} = -\mathrm{d}^2V(x)/\mathrm{d}x^2|_{x=x_*}$. El valor de σ es tanto mayor cuanto mayor sea la curvatura del pozo, y la curvatura es mayor en el caso autorregulado (Figura 2.3.D). De hecho, se puede demostrar con facilidad que la razón de estabilidades entre los casos con y sin autorregulación, $\theta \equiv \sigma_{au}/\sigma_{na}$, viene dada por la sencilla fórmula siguiente:

$$\theta = 2 - \frac{\alpha_{na}}{\alpha_{au}}.$$

El valor de esta razón está acotado entre $\theta = 1$ ($\alpha_{au} = \alpha_{na}$, carencia de control autógeno) y $\theta = 2$ ($\alpha_{au} \gg \alpha_{na}$, fuerte control autógeno). Aunque la ganancia en estabilidad que se da en los sistemas genéticos autorregulados ya había sido demostrada experimentalmente en circuitos sintéticos [74], nosotros queríamos observar dicha ganancia y cuantificar su valor en una red natural como la del SOS. Para esto último, la determinación de los parámetros del sistema a partir de la medida experimental de la AP (ver más abajo) permitieron estimar el valor de la ganancia relativa en estabilidad, que estaba cerca ($\theta \simeq 1.8$) del máximo teórico.

Finalmente, la estabilidad frente a fluctuaciones fue probada infligiendo sobre ambos sistemas una pequeña dosis de daño (que podemos considerar como una perturbación externa sobre el estado de equilibrio de la partícula situada en el fondo del pozo). Sólo en el sistema no autorregulado se pudo apreciar el efecto de esta perturbación, como se muestra en el inserto de la Figura 2.3.D.

2.3.2. Tiempos de recuperación

Al estudiar los tiempos de respuesta del sistema SOS comprobamos cómo los detalles del funcionamiento del sistema son capaces de matizar las propiedades dinámicas que se le podrían atribuir al mismo al extrapolar las conclusiones de estudios genéricos de la autorregulación como los del Capítulo 1. Efectivamente, a partir de los resultados que llevan a afirmar que la presencia del control autógeno proporciona una capacidad de respuesta más rápida [8, 53] podríamos esperar que el sistema natural presentara tiempos de respuesta más rápidos respecto al no autorregulado y que esto fuera así tanto en la trayectoria que va desde el estado no-inducido al inducido como cuando se recorre el camino inverso.

Para probar si esto es así en el caso del sistema SOS cuantificamos la AP de *lexA* que precede al gen *gfp* en ambos circuitos tras someter los cultivos a radiación UV. El proceso se repitió bajo dos regímenes de alta y baja dosis de radiación, tras lo cual obtuvimos las cuatro combinaciones estirpe/dosis que aparecen en la Figura 2.4.

Para empezar, la transición desde el estado sin inducir al inducido es muy similar en todos los casos y, por tanto, parece independiente del tipo de circuito y de la dosis de daño (teniendo en cuenta que las curvas están normalizadas por

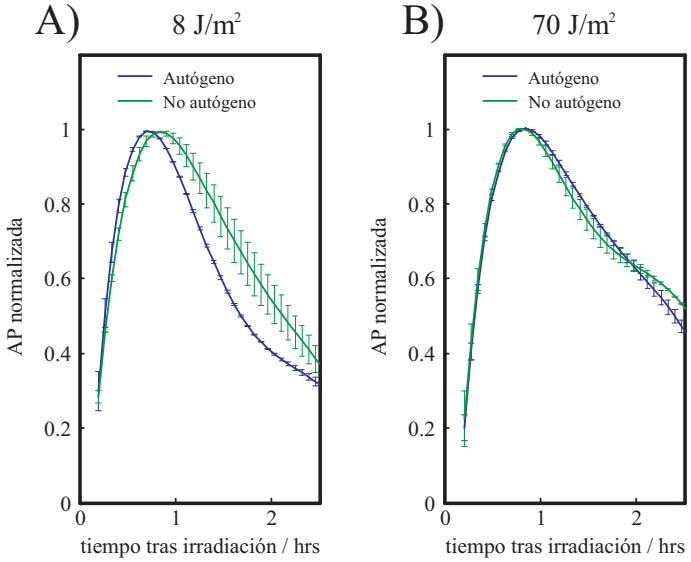


Figura 2.4: *Dinámica de la respuesta SOS tras un proceso de irradiación UV en términos de la AP normalizada. A) AP de los circuitos autógeno y no autógeno bajo un régimen de dosis UV baja (8 J/m^2). B) Ídem para un régimen de dosis UV alta (70 J/m^2).*

su máximo). La dinámica de esta transición está probablemente dominada por el rápido proceso de autodigestión de LexA mediado por RecA* tras el daño [75]. Estaríamos entonces ante una dinámica tan rápida que los diferentes tiempos de respuesta transcripcionales no se hacen notar.

Sin embargo, en el proceso de vuelta al estado no inducido el control autógeno sí que tiene la ocasión de hacer valer su relativa rapidez de respuesta, si bien sólo ante la dosis baja de daño. En estas circunstancias podemos vislumbrar un escenario en el que, tras alcanzar el estado de inducción, la reparación del daño en el ADN se produce con rapidez y el sistema se encuentra de repente con su concentración de LexA en desequilibrio y por debajo de la del estacionario; siendo esto así, el mecanismo compensatorio de la autorregulación –recordemos que si $x < x_*$ la actividad del promotor autorregulado es mayor que la del circuito con AP constitutiva– sería capaz de devolver el sistema al estado de partida con mayor rapidez (Figura 2.4.A). En cambio, a altas dosis de daño los tiempos de reparación del ADN pueden ser lo suficientemente lentos como para que los dos sistemas vayan alcanzando simultáneamente los estados cuasi-estacionarios sucesivos de inducción parcial que se irían produciendo desde el estado de máxima inducción hasta el no-inducido (Figura 2.4.B).

En conclusión, la extrapolación a un sistema natural de todo resultado dinámico –sea teórico o experimental– que esté basado sólo en conexiones transcripcionales puede verse fuertemente matizada por la existencia en el sistema de procesos

cuyas escalas temporales terminen imponiéndose (por rápidas o por lentas) a las propias de la transcripción. Los tiempos asociados a la (rápida) autodigestión de LexA o al (lento) proceso de reparación bajo altas dosis de radiación ejemplifican el tipo de condicionantes dinámicos que podemos encontrar.

2.3.3. Proporcionalidad de la respuesta. Dinámica del sistema SOS con daño en el ADN

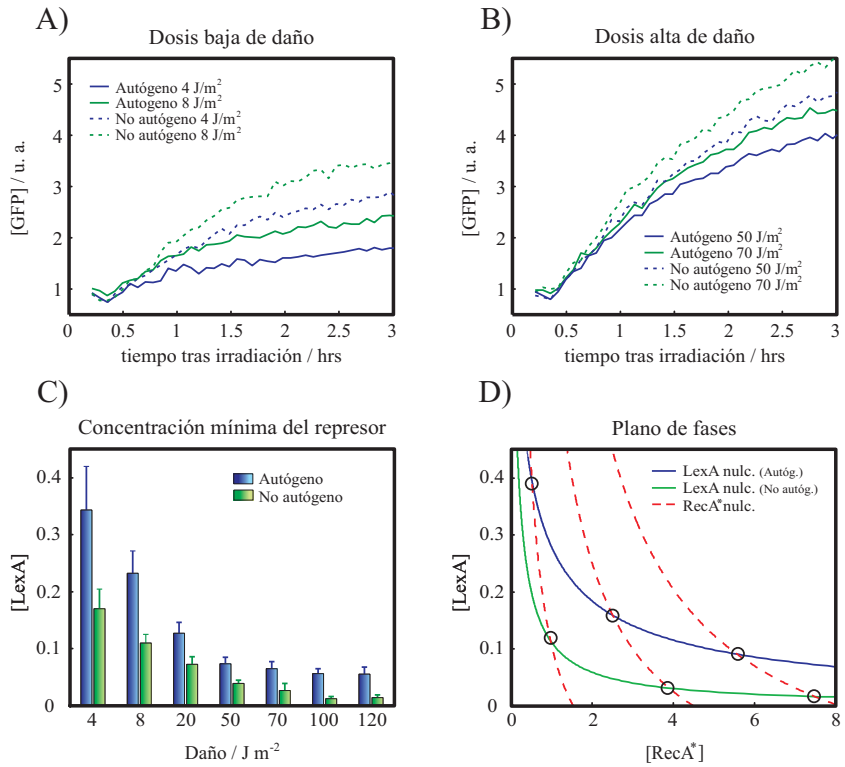


Figura 2.5: *Respuesta SOS ante un amplio rango de dosis UV. A y B) Dinámica del SOS (GFP) en las estirpes portadoras del circuito autorregulado y sin autorregular en condiciones de baja (A) y alta (B) dosis de daño. C) Concentración mínima de LexA en función de la dosis UV para los dos circuitos. D) Plano de fases [LexA]–[RecA*]. Los círculos denotan los estados de equilibrio. La sobrerrespuesta del sistema no-autógeno se refleja en unos estados de equilibrio que implican muy bajas concentraciones del represor LexA. Los valores asignados a los parámetros del modelo pueden consultarse en el artículo del Apéndice B.*

De nuestras medidas se desprende que la autorregulación dota al sistema de una capacidad de respuesta proporcionalmente graduada a los niveles de daño infligidos. En contraste, el control no autógeno sobreinduce la expresión en relación al autorregulado, dándose las mayores diferencias para bajas dosis de daño

(Figura 2.5.A-B). La magnitud de la respuesta está asociada a la mínima concentración de LexA (y la consiguiente máxima AP de la GFP) que cada uno de los sistemas es capaz de alcanzar. La Figura 2.5.C muestra cómo cambia este mínimo bajo las distintas dosis de UV. Como ya hemos reiterado, el mecanismo de la autorregulación es capaz de amortiguar de manera más eficiente la caída en la concentración de LexA, puesto que en ese caso el circuito autorregulado aumenta la tasa de producción de este regulador.

La descripción cualitativa anterior la hemos acompañado de un sencillo modelo matemático que, coherente con estas medidas, nos ofrece una panorámica global de la dinámica del sistema. Se trata de nuevo de un modelo bidimensional como (2.1), aunque la variable y y su tasa de degradación correspondiente δ_y se refieren esta vez a la versión activada de RecA, RecA*. Aún así, los parámetros α_y y k_y siguen siendo, como en (2.1), la AP en ausencia de represor y la constante de disociación de la inactiva RecA, respectivamente[6]. Esto se debe a que asumimos que la transición de RecA a RecA* está en equilibrio, con lo que la AP de esta última se vuelve proporcional a la dinámica de la primera. Finalmente, la diferencia esencial que hace que la inferencia de la funcionalidad del sistema en presencia de daño requiera ir más allá de las conexiones transcripcionales del motivo SIM (ver más arriba), estriba en la existencia del término adicional de la autodegradación de LexA mediada por RecA*[7]. Este término es proporcional a la concentración de ambas proteínas y a la tasa correspondiente δ_{xy}; su inclusión provoca además que, a diferencia de lo que ocurría con (2.1), ahora tengamos un sistema con dos ecuaciones totalmente acopladas,

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t} &= \Pi(x) - \delta_x x - \delta_{xy} xy, \\ \frac{\mathrm{d}y}{\mathrm{d}t} &= \frac{\chi\,\alpha_y}{1+\frac{x}{k_y}} - \delta_y y, \end{aligned} \tag{2.5}$$

donde $\Pi(x)$ es la AP de LexA que depende del circuito de que se trate: $\Pi_{au}(x) = \alpha_{au}/(1 + x/k_x)$, $\Pi_{na}(x) = \alpha_{na}$, con la misma relación (2.2) entre α_{au} y α_{na} con la que en ambos circuitos se obtiene una misma concentración estacionaria en ausencia de daño. χ es la fracción de proteínas RecA que están activadas y su rango de valores abarca desde 0 (ausencia de daño) a 1 (virtualmente todo RecA está activado en forma de RecA*). Este parámetro lo imponemos externamente para modelizar situaciones que cubran muy diferentes dosis de daño. Nótese que al fijar un valor constante para χ no estamos considerando el proceso de reparación

[6]Para mantener la consistencia de la notación matemática a lo largo de este capítulo existen pequeñas diferencias con respecto a la que se sigue en el artículo del Apéndice B. Así, las constantes k_x, k_y y α_y son allí k, k' y β, respectivamente.

[7]Ya hemos visto, al tratar los tiempos de respuesta, cómo el proceso postraduccional de degradación de LexA matiza los resultados que derivan de estudios basados exclusivamente en conexiones transcripcionales.

–que haría decrecer en el tiempo el valor inicial de este parámetro hasta anularlo. Por ello, esta descripción se corresponde con la etapa inicial en la que la proteólisis de LexA domina la dinámica del sistema, etapa que concluye con la concentración de esta proteína en un valor mínimo –que coincidiría, en cada situación de estirpe y daño, con el punto álgido de la curva de AP correspondiente. Dicho valor mínimo constituye precisamente la solución estacionaria del sistema (2.5).

El plano de fases representado en la Figura 2.5.D muestra las nulclinas del sistema de ecuaciones (2.5) con y sin autorregulación, y para tres valores del parámetro de daño χ. Los dos circuitos autorregulado y sin autorregular comparten la ecuación para y y, con ello, la nulclina para RecA* ($\mathrm{d}y/\mathrm{d}t = 0$),

$$y(x) = \frac{1}{\delta_y} \frac{\chi\, \alpha_y}{1 + \frac{x}{k_y}}. \tag{2.6}$$

La ecuación (2.6) constituye una familia χ–paramétrica de curvas cuya localización se desplaza hacia valores más pequeños de RecA* conforme el daño decrece. La nulclina de RecA* llega a colapsar con el eje x en ausencia de daño (nótese que los ejes están girados en la Figura 2.5.C-D). En cambio, las nulclinas para LexA ($\mathrm{d}x/\mathrm{d}t = 0$) son diferentes en cada circuito,

$$\begin{aligned} y(x)_{au} &= \frac{1}{\delta_{xy}} \left[\frac{\alpha_{au}}{x(1 + x/k)} - \delta_x \right], \\ y(x)_{na} &= \frac{1}{\delta_{xy}} \left(\frac{\alpha_{na}}{x} - \delta_x \right). \end{aligned}$$

En ausencia de daño ($y = 0$) ambas cortan al eje x en la solución estacionaria x_* –en ese caso los primeros términos a la derecha de cada ecuación se igualan debido a (2.2). Los puntos de cortes de las nulclinas para x e y en la Figura 2.5.D marcan las soluciones estacionarias que dan el valor mínimo que alcanza en cada caso la concentración de LexA. Como se ve al comparar el par de soluciones que, correspondientes a uno y otro circuito, se da en cada una de las condiciones de daño, la máxima diferencia absoluta se produce para bajos niveles de radiación. Y aunque el mínimo que alcanza el sistema sin autorregular está siempre por debajo del correspondiente al sistema con control autógeno (con la correspondiente sobreinducción de la respuesta), la diferencia absoluta decrece según aumenta el daño, en concordancia con los resultados experimentales de la Figura 2.5.C.

2.3.4. Consecuencias para el crecimiento

La sobrerrespuesta del circuito no autorregulado implica la sobreproducción de toda la batería de proteínas reguladas por LexA en el sistema SOS, con el consiguiente coste energético extra. Además, el producto de dos de los genes SOS, *umuD* y *umuC*, puede ralentizar el ciclo celular [76]. Por ello, nos propusimos

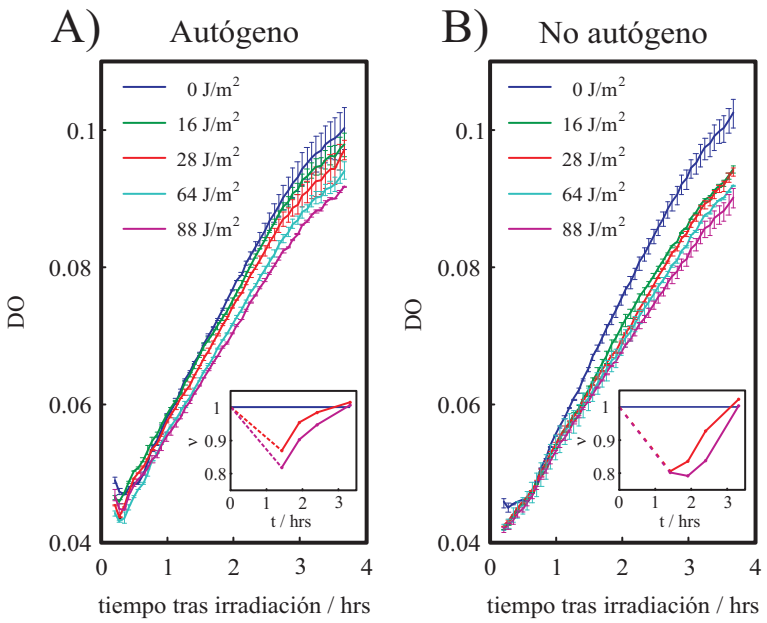


Figura 2.6: *Curvas de crecimiento (DO) tras el proceso de radiación UV de las estirpes portadoras del circuito autorregulado (A) y sin autorregular (B). Insertos: evolución de la tasa de crecimiento ν.*

abordar la cuestión de si la diferencia en la dinámica de la respuesta, derivada de la presencia o ausencia de autorregulación, tiene su reflejo sobre las tasas de crecimiento de las estirpes correspondientes. En la Figura 2.6.D está trazada la evolución de la DO de las estirpes portadoras de los circuitos natural y sintético. En cada una de las estirpes por separado la exposición a niveles progresivamente más elevados de radiación UV se corresponde con un descenso en la tasa de crecimiento; es decir, que toda respuesta tiene un coste. Sin embargo, ante una misma dosis de radiación, la sobrerrespuesta asociada al circuito sin control autógeno implica siempre tasas de crecimiento menores a las de la estirpe con el circuito autorregulado.

En conjunto, éste y los resultados anteriores apuntan a que la presencia de la autorregulación de LexA tiene por objeto obtener un ajuste fino de la respuesta del sistema, de manera que éste no se sobreinduzca innecesariamente ante niveles bajos de daño. Probablemente, la presencia mayoritaria de la autorregulación en otras redes que exhiben una arquitectura tipo SIM [9] obedezca a la misma razón: evitar una respuesta excesiva cuyo coste se vería amplificado por toda la batería de genes regulados.

Capítulo 3

La autorregulación en el ensamblaje de los motivos de la red de transcripción de *Escherichia coli*

En este capítulo se presentan los resultados del siguiente artículo, incluido en el Apéndice C:
Camas FM, Poyatos JF (2008) What Determines the Assembly of Transcriptional Network Motifs in *Escherichia coli?* PLoS ONE 3(11): e3657. doi:10.1371/journal.pone.0003657
Con el propósito de incidir en los aspectos relacionados con la autorregulación, en algunos momentos no se sigue estrictamente el orden de presentación de los resultados tal y como aparecen en la publicación. Para abreviar, el artículo y su documento suplementario, también incluido en el Apéndice C, serán referenciados a lo largo de este capítulo como Artículo *y* Suplemento*, respectivamente.*

3.1. La identificación motivo/módulo

En el Capítulo 2 hemos estudiado un motivo SIM que se comporta como un módulo funcional debido a que las especies moleculares codificadas en los genes-nodos de esta subred desempeñan una función con un alto grado de autonomía respecto al resto de funciones celulares [5]. Además, allí mostrábamos cómo la autorregulación del factor de transcripción (TF) maestro de este SIM incide en la dinámica del sistema. Si esta identificación motivo-módulo fuese extrapolable al resto de los motivos de la red de transcripción, el panorama se presentaría muy prometedor de cara a la racionalización del funcionamiento dinámico del conjun-

to de la red, una racionalización que tendría ese carácter de abajo a arriba que implican las arquitecturas modulares. Por otro lado, la autorregulación aparece vinculada tanto a los reguladores maestros de los SIM como a los elementos Y de los FFLs [10] (Figura 3.2.E) y, por tanto, la autorregulación podría estar condicionando de nuevo, como en el caso del SOS, la dinámica de estos supuestos motivos-módulos.

La existencia de una funcionalidad celular implica un proceso subyacente de selección adaptativa. Por ello, si se considera que la significatividad estadística de los motivos supone un indicio de este tipo de selección, entonces habría que atribuirles una función; en concreto, los defensores de esta línea argumental asignan a los motivos tareas específicas de procesamiento de información [10]. Muchos de los motivos sólo involucran a unos pocos nodos de la red (Figura 2.1) y, por ello, salvo en el caso del SIM, el número de especies moleculares implicadas es demasiado bajo como para dar cuenta de las funciones celulares más complejas; esto impide asociar directamente los motivos a los módulos funcionales. No obstante, la frecuente aparición de autorregulaciones, FFLs y bi-fanes en forma de agregados ha dado pie a la proposición de estos agregados como entidades funcionales de mayor envergadura, de las cuales los motivos serían las unidades básicas de *computación* [17–19]. De acuerdo con este planteamiento, la tarea que cada clase de motivo lleva a cabo viene determinada por la propia arquitectura del motivo, con lo que, en consecuencia, la comprobación experimental de que esto es así en representantes de cada una de las clases debe iluminar la dinámica del resto de sistemas en los que los motivos aparecen [10].

Ya vimos en el capítulo precedente las críticas a la hipótesis de que la estructura transcripcional de los motivos determina su función; y cómo parte de las mismas eran aplicables al propio sistema SOS debido a la existencia de una relación postraduccional que en gran medida condiciona la dinámica del mismo. A estos argumentos añadamos ahora los que, en oposición al escenario adaptativo, propugnan que los motivos surgen debido a la acción de fuerzas neutrales poblacionales como la deriva genética [15] o como resultado de los mecanismos intrínsecos de la evolución genómica –recombinaciones y duplicaciones a pequeña y gran escala [77–79]. Se ha llegado incluso a plantear que la señal estadística de los motivos emerge simplemente como consecuencia de la elección de un modelo nulo de red inadecuado [15, 64, 65]. Además, los defensores de este escenario neutral ven reforzadas sus tesis por la carencia de una señal clara de conservación de los motivos concretos entre distintas especies [80].

A todo esto hay que sumar la cuestión del grado de significatividad estadística de *cada uno* de los casos concretos de motivos. La descripción original de los motivos de red descansaba en el argumento de la sobrerrepresentación global de dichos motivos en la red natural frente a modelos *nulos* de redes aleatorias [9, 61]. Pero, incluso asumiendo que cada una de las regulaciones presentes en la red natural ha emergido de manera adaptativa [11, 80–83], no se puede deducir de

manera inmediata que cada uno de los motivos existentes en la misma ha sido seleccionado como un ente funcional. Podemos pensar en un escenario en el que esto podría ser así sólo en algunos casos –que serían entonces responsables en gran parte de la señal global que dio pie a la descripción de los motivos– pero que, en cambio, pueden existir otros ejemplos de motivos que no sean más que un efecto colateral del establecimiento de las regulaciones que los constituyen[1].

En la siguiente sección se consideran los detalles que subyacen a la señal global de los motivos SIM. Los casos del FFL y del bi-fan serán tratados más adelante.

3.2. El sistema SOS no es cualquier SIM

Recordemos que la definición del SIM exige que los operones regulados por el TF maestro lo sean de manera exclusiva, es decir, que sobre aquéllos no exista más regulación transcripcional que la que le viene de este regulador. Esto hace que, por definición, el SIM sea un motivo bastante aislado del resto de la red de transcripción y, por tanto, un motivo al que es más fácil atribuir una funcionalidad autónoma. Pero, aún así, veremos que este aislamiento no es significativo en muchos casos.

De acuerdo con la definición original de SIM [9] –regulación en exclusiva (y bajo el mismo signo, activación o represión) de tres o más operones–, en un modelo neutro de red aleatoria se pueden formar tantos (si no más) SIMs que en la red natural (red de *E. coli*: 27, promedio redes aleatorias: 30.5, $p = 0.11$)[2]. Esto es debido a que la regulación en exclusiva de unos pocos operones se obtiene con frecuencia en el recableado aleatorio de una red con 681 nodos, 1109 conexiones no autorregulatorias y 135 TFs[3]. Sin embargo, el número medio de operones regulados por los TFs maestros es significativamente mayor en el caso natural (10.3 frente a 8.4 en las redes aleatorias, $p = 0.0032$). Esto nos devuelve la conclusión –ya planteada en el artículo original que por primera vez describía los motivos transcripcionales [9]– de que sólo los SIM que implican a muchos operones regulados son realmente motivos de red. No obstante, se podrían aún arrojar dudas sobre la autonomía funcional de estos SIM significativos si el conjunto de operones regulados en exclusividad supusiese sólo una fracción pequeña del conjunto de operones que están bajo el control del regulador maestro del SIM.

[1] Nótese que todas estas consideraciones son tenidas en cuentas al elaborar un modelo nulo de red transcripcional en el que las conexiones de la red natural son recableadas de una manera aleatoria pero manteniendo el número de nodos, las conexiones y el signo (de activación o represión) de éstas (*Artículo*, Materiales y métodos).

[2] Los detalles sobre el cálculo de p-valores y otros procedimientos estadísticos pueden consultarse en *Artículo* y *Suplemento*, sección 2.

[3] Todos los números que aparecen en este capítulo concernientes a la red de regulación transcripcional de *Escherichia coli* corresponden a la versión de la misma que hemos ensamblado a partir de los datos de RegulonDB versión 5.6 [84] (*Artículo*, Materiales y métodos).

Esta última consideración ha sido tenida en cuenta cuando hemos abordado el análisis de la significatividad estadística de cada uno de los casos de motivos SIM que aparecen en la red. Así, para llevar a cabo este análisis introdujimos una magnitud que daba cuenta de la tendencia de un TF dado a formar un motivo SIM. A esta magnitud la denominamos la *SIMness*[4] ($\mathcal{S}$, $0 \leq \mathcal{S} \leq 1$) del TF y se define como la proporción entre el número de operones que el TF regula con exclusividad bajo el mismo signo –de activación o de represión– y el tamaño del regulón de ese signo[5] (Figura 3.1).

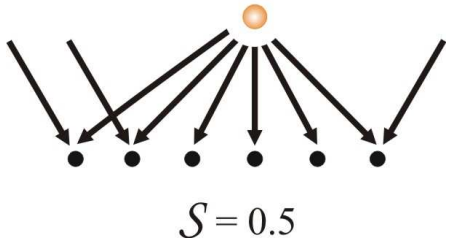


Figura 3.1: *Cálculo de la SIMness. El regulón del TF (en naranja) abarca 6 operones (nodos en negro). De éstos, sólo 3 son regulados de una manera exclusiva, puesto que sobre los demás operones concurren regulaciones adicionales. Por tanto, en este caso el valor de la SIMness del regulador maestro es* $\mathcal{S} = 0.5$.

Esta magnitud se puede comparar con la que se obtendría en redes aleatorias, lo que permite evaluar caso por caso la significatividad de la SIMness de cada TF. La Tabla 3.1 muestra el número limitado de casos en los que la SIMness se reveló significativamente grande y que corresponden a TFs de regulones grandes que son regulados en buena parte de manera exclusiva. También aparecen dos casos con SIMness significativamente pequeña, CRP e IHF. Aunque se trata de nuevo de TFs con regulones grandes, esta vez la regulación de muchos de los operones es compartida con otros TFs distintos al maestro. Nótese que en el caso de CRP, el regulador global por antonomasia de *E. coli*, la pequeña fracción de operones regulados en exclusiva (18 en términos absolutos) da lugar a la formación de un motivo SIM que se consideraría significativo bajo la definición basada en la señal estadística global.

Ya en esta Tabla vemos que el sistema SOS no es un ejemplo más entre los distintos motivos SIM de *E. coli* puesto que LexA, su regulador maestro, es el que tiene la más alta SIMness. Si unimos a esto el alto grado de aislamiento propio de los SIM, resulta muy fácil hacer en este caso la asociación motivo/entidad-

[4]Que en español se podría traducir por algo así como "SIMneidad", la cualidad de SIM. En cualquier caso, mantendremos el término anglosajón usado en el *Artículo*.

[5]Para ser consistentes con la definición de SIM, sólo definimos la SIMness para aquellos TFs en los que este tipo de regulón "monocolor" implicara a tres o más operones. Las interacciones duales no fueron consideradas.

TF	R^+	$\mathcal{S}^+$	$\mathcal{S}_r^+$	Z-score*
RpoE	51	0.78	0.33	7.34
CRP	117	0.15	0.38	-5.98
Fis	41	0.68	0.33	5.10
RpoH	25	0.68	0.31	4.07
IHF	28	0.07	0.33	-3.01
TF	R^-	$\mathcal{S}^-$	$\mathcal{S}_r^-$	Z-score*
LexA	19	0.89	0.31	5.65
Fur	28	0.57	0.31	3.20

Tabla 3.1: *Factores de transcripción con SIMness significativamente grande o pequeña.* R^i*: tamaño del regulón,* $\mathcal{S}^i$*: SIMness,* $\mathcal{S}_r^i$*: promedio de la SIMness en redes aleatorias (*r, random*), i: regulación positiva o negativa.*
**Sólo se consideraron los casos en los que el p-valor era mayor que 0.05 tras hacer el ajuste por comparación múltiple [85].*

funcional-autónoma, así como explicar el papel de la autorregulación en el esquema de funcionamiento del motivo.

Finalmente, la dificultad que existe para establecer un vínculo entre significatividad estadística y funcionalidad queda patente al considerar el siguiente ejemplo de motivo SIM, que exhibe la segunda pero no la primera de estas características. Se trata del sistema de la biosíntesis de la arginina, en el cual el represor ArgR regula de manera exclusiva a 5 de los 9 operones de los que consta su regulón. La arquitectura SIM constituida con esos 5 operones exhibe un comportamiento dinámico tan interesante como un patrón de activación de la transcripción de los operones regulados en un orden temporal que se corresponde con el de la posición en la ruta metabólica de las enzimas codificadas en éstos [86], lo que se ha considerado como un muestra paradigmática de las propiedades de computación de los SIMs [10]. No obstante, aun cuando nadie duda de la funcionalidad autónoma del sistema, en este caso ni la arquitectura SIM involucra a un número grande de operones ni el valor de la SIMness ($\mathcal{S} = 0.56$) de ArgR es estadísticamente significativo.

3.3. Distribución de la autorregulación en la red de transcripción

Aproximadamente el 56 % de los TFs de la red de transcripción de *E. coli* están autorregulados[6]. Para empezar con el estudio de la integración de la autorregulación en motivos más complejos, vamos a considerar la relación de la presencia de la autorregulación de un nodo de la red con las regulaciones que llegan o salen del mismo. A este respecto se pueden plantear dos cuestiones complementarias (Figura 3.2.A): i) ¿suele actuar la autorregulación en solitario o en combinación con otras interacciones transcripcionales entrantes? y ii) ¿depende la presencia del control autógeno en un TF del número de operones que éste regula?

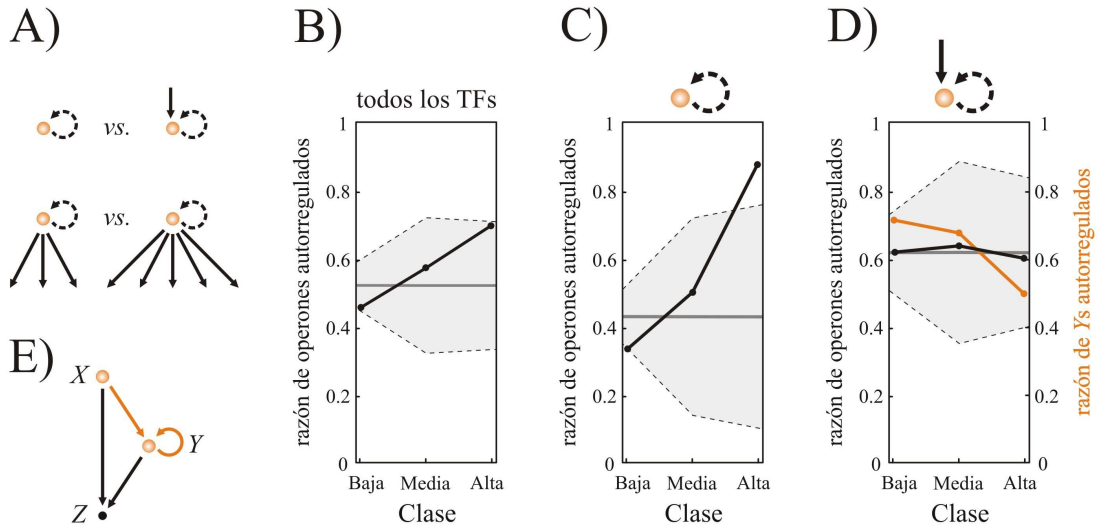


Figura 3.2: *Distribución de la autorregulación. A) Evaluamos frecuencia de autorregulaciones en función de i) presencia/ausencia de regulación externa (arriba) y ii) tamaño del regulón (abajo). B) Frecuencia de autorregulaciones en las tres clases de conectividad. Clases de conectividad: baja (de 1 a 4 operones regulados), media (de 5 a 9) y alta (10 o más, que son los nodos principales o* hubs *de la red). Se superpone el comportamiento nulo que se obtiene mediante muestreos aleatorios (media, linea gris; ±2 desviaciones estándar, área sombreada; véase* Suplemento, *sección 2). Líneas entre puntos para ayudar a la visualización. C-D) Como B) pero distinguiendo entre aquellos TFs sin (C) y con regulación externa (D). En D) se muestra también la proporción de elementos Y autorregulados (en naranja). E) TF autorregulado con regulación externa como parte de un FFL.*

[6]La sustancial presencia de TFs autorregulados ya fue documentada en el ensamblaje previo de la red de transcripción sobre el que se describieron originalmente los motivos [9, 42]. En la sección 1 del *Suplemento* se realiza una comparación exhaustiva entre este ensamblaje, que involucraba un menor número de nodos y conexiones, y el nuestro propio.

Obsérvese que la primera pregunta equivale a reducir la estructura multicapa de la red de TFs [19, 42, 87, 88] a sólo dos niveles: el superior estaría formado por aquellos TFs que no experimentan ninguna regulación transcripcional *externa*[7]; en este caso, las autorregulaciones actuarían de manera aislada. En cambio, en el nivel inferior tendríamos el resto de TFs en los que la autorregulación siempre se presentaría combinada con alguna regulación transcripcional adicional. Pues bien, el nivel superior está en general menos autorregulado (27 de 63 TFs están autorregulados, 43 %) que el nivel inferior (37/60, esto es, un 62 % de los TFs, $p = 0.03$). Por tanto, considerados en su conjunto, existe en promedio una mayor incidencia de la autorregulación entre los factores de transcripción que tienen regulación externa que entre los que no la tienen (Figura 3.2.C-D).

Sin embargo, y con esto abordamos la segunda de las cuestiones planteadas, las cosas se matizan al considerar las conexiones salientes: mientras que entre los TFs del nivel inferior la incidencia de la autorregulación parece desacoplada del tamaño de los regulones (Figura 3.2.D, línea negra), en el nivel superior hay una tendencia muy acusada a que la autorregulación crezca con la conectividad saliente (Figura 3.2.C). De hecho, los nodos con grandes regulones (*hubs*[8]) del primer nivel llegan a suponer una excepción a la baja incidencia de la autorregulación entre los TFs carentes de regulación externa (7 de los 8 hubs que hay en el primer nivel están autorregulados). En general, estos TFs son capaces de provocar, ante el estímulo adecuado, grandes cambios en la fisiología celular. Por tanto, podríamos plantear la hipótesis de que, en ausencia de regulación transcripcional externa que controle la expresión de este tipo de TF, la autorregulación proporcionaría dicho control[9].

Como mencionamos más arriba, se ha propuesto la asociación de la autorregulación con el elemento central de los FFLs (Y en la Figura 3.2.E). Dado que, por definición, este elemento Y tiene regulación externa (ejercida por X), el enriquecimiento de las capas bajas de la red de transcripción con elementos autorregulados podría estar relacionada con esta manera de integrarse la autorregulación en los FFLs[10]. Esta posibilidad la abordaremos en la sección siguiente.

[7]Es decir, proveniente de otro TF distinto.

[8]Vamos a clasificar los operones en tres clases de alta, media y baja conectividad saliente (tamaño del regulón). Usaremos también el término anglosajón *hub* (a partir de ahora, sin cursivas) para referirnos al primer grupo, que incluye a los TFs que regulan a diez o más operones (véase el pie de la Figura 3.2).

[9]La regulación ejercida por cada uno de estos siete TFs y, por ende, el propio control autorregulatorio están condicionados siempre por la presencia de algún tipo de cofactor que propaga de señal de entrada del sistema [89]. Ya hemos visto que en el caso de LexA (que es un hub de primer nivel) la señal de daño en el ADN es propagada a través de RecA* (Capítulo 2). En otros casos el papel de cofactor lo ejerce una molécula orgánica no proteínica: cAMP para CRP, y los aminoácidos L-arginina y L-leucina para ArgR y Lrp, respectivamente. Para PhoP y CpxR, miembros de sistemas de dos componentes, las señales de estrés se propagan mediante la fosforilación de estos TFs. Finalmente, la regulación ejercida por FNR es desactivada por vía de la oxidación de su grupo [2Fe-2S].

[10]Si esta integración se realizara de una manera estadísticamente significativa, de por sí con-

3.4. Integración de la autorregulación en los FFLs

Si, a partir de la clasificación de los 232 FFLs de la red en tres grupos correspondientes a la conectividad[11] del elemento Y, se calcula el porcentaje de FFLs en los que este elemento exhibe autorregulación, se obtiene que este porcentaje decae con la conectividad –desde un 71 % de elementos Y autorregulados cuando el regulón de Y es pequeño hasta sólo un 50 % entre los Y de alta conectividad. Así, la Figura 3.2.D muestra el manifiesto contraste entre el comportamiento indiferente a la conectividad que muestra la distribución de las autorregulaciones entre los TFs con regulación externa, y el sesgo que la asociación de la autorregulación con los elementos Y de los FFLs tiene hacia los TFs de baja conectividad.

Sin embargo, esta comparación sólo supone una primera intuición de que la relación de la autorregulación y los FFLs podría limitarse únicamente al grupo de baja conectividad. Ello es debido a que, dado que los FFLs se muestran muchas veces en forma de agregados que comparten un mismo elemento Y [17–19], en el recuento anterior se están considerando TFs de manera redundante. Por ello, los porcentajes de elementos Y autorregulados no son rigurosamente comparables con los obtenidos sobre la población de TFs con regulación externa.

Para evitar estas redundancias necesitábamos una medida que diese cuenta a la vez de todos los FFLs que se establecen sobre un mismo elemento Y. Por ello, definimos una magnitud, la *FFLness* ($\mathcal{F}$, $0 \leq \mathcal{F} \leq 1$), que sería aplicable a cada TF de las capas inferiores de la red que regulase al menos a un operón (sin contar al propio que lo codifica). Esta medida se define pues como la proporción entre el número de FFLs que tienen un mismo TF como elemento Y y el máximo número de FFLs que dicho TF podría constituir potencialmente, máximo dado por el producto del número de regulaciones entrantes sobre el TF y el tamaño de su regulón (Figura 3.3.A)[12].

La Figura 3.3.B-D (líneas negras) muestra la FFLness promedio para los TFs de cada clase de conectividad. La tendencia general, independientemente de que los TFs estén autorregulados o no, es que $\mathcal{F}$ decrezca mucho con el tamaño del regulón. La FFLness de los hubs -aun permaneciendo dentro de lo significativo- se acerca mucho a los valores del modelo neutro (zonas sombreadas en gris)[13]. Por

tribuiría a la consideración de los FFLs como entes funcionales, puesto que aumentaría la complejidad del motivo, y con ello la improbabilidad del mismo.

[11] Por concisión en el discurso, al referirnos en general a la conectividad de un TF lo hacemos a la conectividad *saliente* (número de operones regulados por el TF).

[12] Nótese el paralelismo de esta definición con la de la SIMness. En ambos casos, se tratan de magnitudes que permiten individualizar la señal estadística global de los motivos. Como con la SIMness, vamos a mantener el término en inglés. Todo intento de traducir "feed-forward-loopness" al español está abocado a la cacofonía.

[13] Sobre la FFLness en el modelo neutro: como se puede ver a partir de su definición y en el ejemplo de la Figura 3.3.A, la FFLness es una cantidad que presenta un alto grado de discretización entre los TFs de la clase de baja conectividad. El número de regulaciones entrantes es

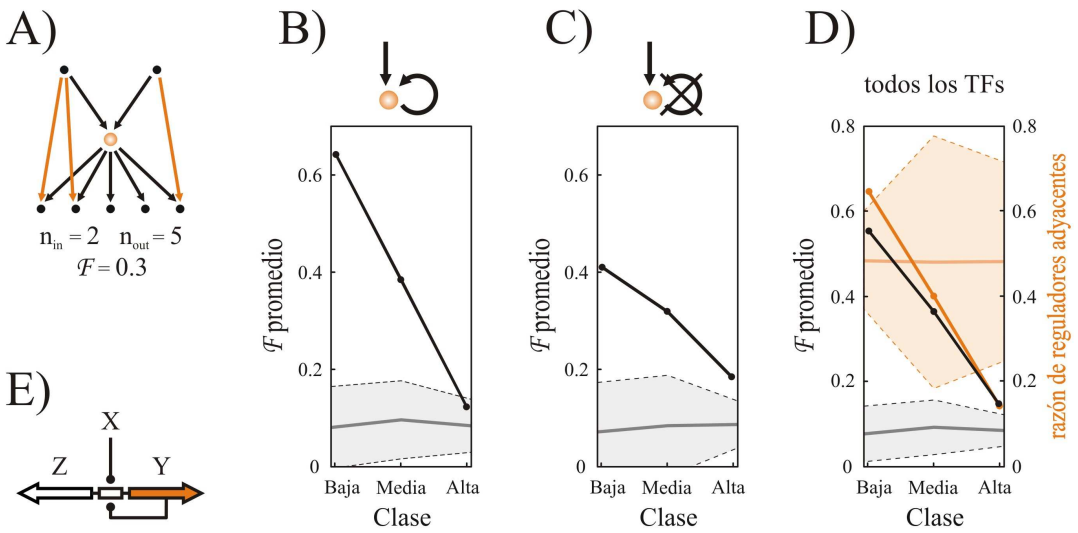


Figura 3.3: *Ensamblaje de FFLs. A) Cálculo de la FFLness: el máximo número de FFLs que pueden ser ensamblados en este caso es* $n_{\text{in}}n_{\text{out}} = 10$*. Al constituirse de facto 3 FFLs (gracias a las conexiones pintadas en naranja), la FFLness del TF central vale* $\mathcal{F} = 0.3$*. B-D) FFLness media como función del tamaño del regulón para aquellos TFs con control externo (clases de conectividad definidas en la Figura 3.2). Mientras que en B) y C) distinguimos entre TFs autorregulados y no autorregulados, respectivamente, en D) se consideran todos en conjunto. Se superpone el comportamiento nulo (media, línea gris;* ± 2 *desviaciones estándar, área sombreada; véase* Suplemento*, sección 2). En D) se muestra también la proporción de TFs –con regulación externa– que regulan a un operón vecino (línea naranja oscuro) y su correspondiente nulo (media, línea naranja claro continua;* ± 2 *desviaciones estándar, área sombreada en naranja). Líneas entre puntos para ayudar a la visualización. E) La acción reguladora sobre la zona intergénica entre los operones divergentes (*Y *y* Z *en el FFL) de los factores de transcripción* X *e* Y*, da lugar a la formación de un FFL que tiene el elemento* Y *autorregulado.*

otro lado, la distribución de la autorregulación en los elementos Y de las distintas clases de conectividad –que obtuvimos con el primer recuento en bruto– se confirma en la comparación (dentro de cada clase) de los promedios de FFLness del subconjunto de los autorregulados (Figura 3.3.B) frente al de los no-autorregulados (Figura 3.3.C): el exceso de autorregulación en los Y de baja conectividad se co-

casi siempre muy bajo en toda la red de transcripción. Así, cuando el número de regulaciones salientes es igual a 1, la FFLness sólo puede tomar valores extremos, o cero o uno. Pero, como se puede observar, el promedio de $\mathcal{F}$ en el modelo neutro es prácticamente constante a lo largo de las clases de conectividad. El efecto discreto en las bajas conectividades se diluye cuando i) en una randomización se promedia la FFLness dentro de una clase y ii) se promedia sobre el resultado de muchas randomizaciones (que es lo que finalmente aparece en la Figura). Así, de lo discreto apenas queda un reflejo en el algo mayor valor de la desviación estándar (zonas sombreadas) que el modelo neutro muestra en la clase de baja conectividad.

rresponde con el mayor promedio de $\mathcal{F}$ entre los autorregulados de esa clase. La diferencia mengua –aunque todavía decantándose hacia los autorregulados– en el caso de los de media conectividad para finalmente invertirse en los de alta[14].

Por tanto, parece que la señal estadística global de los FFLs está soportada en gran medida por los FFLs con Y de baja conectividad, y que, si bien el promedio de $\mathcal{F}$ no abandona en ningún caso la zona de lo estadísticamente significativo, la contribución de las otras dos clases a esta señal es menor (sobre todo la que viene de los hubs). Dejaremos para más adelante la cuestión de si esto pone en tela de juicio la consideración de muchos de los FFLs con Y de alta conectividad como meros efectos colaterales del establecimiento de las conexiones de la red transcripcional y no como sujeto de selección adaptativa; así, de momento quedémonos con el hecho de que la autorregulación aparece vinculada a los FFLs más significativos, contribuyendo a la consideración de estos casos como unidades funcionales.

No obstante, esta imagen última no es completa: en primer lugar el vínculo entre autorregulación y alta FFLness no es estadísticamente significativo (TFs autoregulados: $\mathcal{F} = 0.64$; no autoregulados: $\mathcal{F} = 0.41$, $p = 0.12$, Figura 3.3.B–C, clase de baja conectividad). En segundo lugar, la autorregulación en el nivel inferior no está sesgada hacia los TF de baja conectividad, como cabría esperar si la autorregulación de un TF de los niveles inferiores de la red estuviese fuertemente condicionada por la pertenencia de este TF a un FFL. Por tanto, o este vínculo no es estrecho o, si lo es, la presencia de la autorregulación en los TF de más alta conectividad obedece a otras razones distintas a su integración en la arquitectura del FFL. Por último, recordemos que la consideración de los motivos como entidades de procesamiento específico de información descansa en la aceptación de la equivalencia entre selección adaptativa y significatividad estadística –venga ésta de recuentos globales de motivos o de magnitudes individualizadas tipo $\mathcal{S}$ o $\mathcal{F}$. Como hemos visto más arriba, esta equivalencia es objeto de un intenso debate.

En la siguiente sección se aborda este último punto, que es el más crítico. Así, trataremos de afianzar el escenario adaptativo para los FFLs con Y de baja conectividad, aportando argumentos que, yendo más allá de los meramente estadísticos, identifiquen en estos FFLs la selección de una funcionalidad biológica. Finalmente, cerraremos el círculo reencontrándonos con la autorregulación, la cual, por la manera en la que se integra en esta funcionalidad, explica su vinculación con los FFLs.

3.4.1. Afianzando el escenario adaptativo en los FFLs con Y de baja conectividad.

Por tanto, antes que nada, y para evitar seguir la pista a un artefacto, nos planteamos si la significatividad de este grupo de FFLs no es debida a su selec-

[14]Nótese que este comportamiento es coherente con el de la línea naranja de la Figura 3.2.D.

ción adaptativa sino a mecanismos neutrales que pudieran estar ocasionando un incremento de la señal de la FFLness para los TF de baja conectividad. Los dos escenarios que hemos considerado que podrían dar lugar a una constitución neutral de estos FFLs tienen que ver con la arquitectura del genoma y con posibles homologías entre los constituyentes del FFL.

Como nuestra intención era asegurar la conexión de la significatividad estadística con la selección adaptativa, adoptamos una postura cauta antes de adjudicar a un FFL la etiqueta de "adaptativo". Por ello, nuestra manera de proceder fue descartar aquellos FFLs cuyo ensamblaje pudiera ser atribuido a estas fuerzas neutrales, siendo generosos a la hora de llenar el saco de los no adaptativos. Aunque esto da lugar probablemente a un notable incremento de los falsos negativos, a cambio se obtiene una razonable seguridad de que los motivos que sobrevivan a la criba son buenos candidatos a haber sido seleccionados de manera adaptativa (pocos falsos positivos)[15].

El proceso de descarte que veremos a continuación se hizo de manera secuencial: primero se filtraron los casos atribuibles a la arquitectura genómica; después, a los supervivientes se les aplicó el criterio de la homología. La clasificación final de los 74 FFLs con Y de baja conectividad obtenida tras este proceso se puede ver en la Figura 3.4.A (clase baja).

Primer escenario neutral: ¿es la alta FFLness una consecuencia de la arquitectura genómica?

La alta FFLness podría deberse al fenómeno de *regulación del vecino* que se da entre los TFs que regulan a muy pocos operones, y que consiste en la localización adyacente en el genoma del operón que codifica al TF y alguno de los operones regulados por éste [25–28]. La Figura 3.3.D (línea naranja) muestra la proporción de TFs que, teniendo regulación externa, regulan al menos a un operón vecino. El comportamiento de esta proporción es el mismo que el de la FFLness –alta para regulones pequeños y decaimiento para los de mayor tamaño– sugiriendo que esta regulación local podría ser un factor subyacente a la alta FFLness. Pero, ¿cómo puede condicionar la arquitectura genómica el valor de $\mathcal{F}$?

La orientación relativa entre el operón que codifica el regulador y el operón adyacentemente regulado que se da con más frecuencia en el genoma de *E. coli* es la disposición divergente [27], la cual es propicia para vincular regulación adyacente y autorregulación [25–28] y, en última instancia, para la constitución de un FFL con el elemento Y autorregulado. Esto se debe a que: i) la regulación del operón adyacente y la autorregulación operan en una misma zona intergénica; ii) cualquier

[15] Ni que decir tiene que esta estrategia se basa en que, a pesar de todo, el número de positivos verdaderos va a ser al final una fracción importante de la población de FFLs considerada, corroborando así la hipótesis de partida de que la señal es significativa por provenir de un proceso de selección adaptativa.

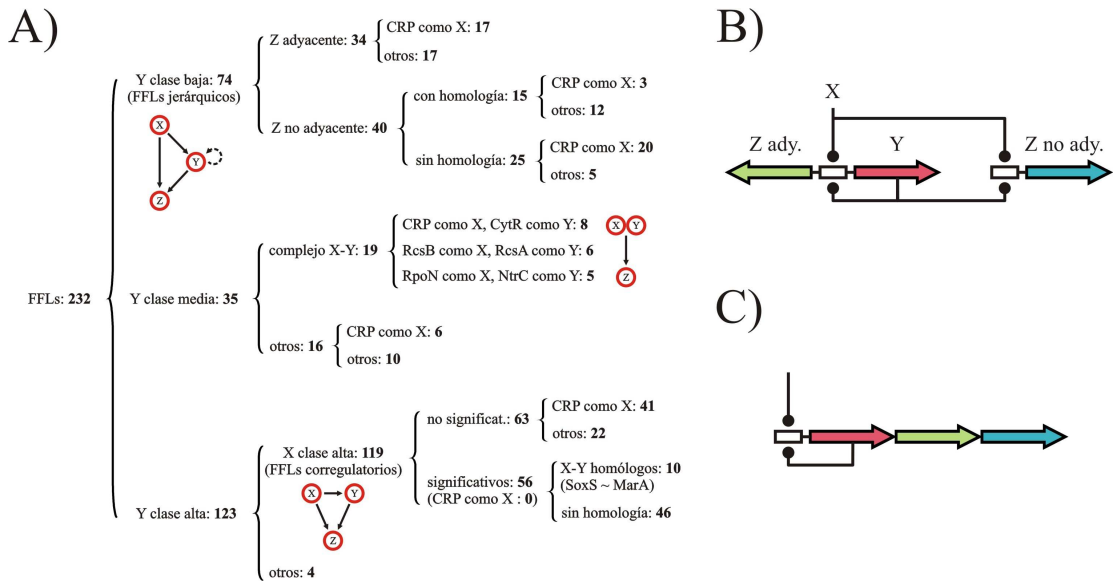


Figura 3.4: *Clasificación de los FFLs. Los 232 FFLs identificados en la red son agrupados en función de la clase de (baja, media o alta) conectividad del elemento Y (véase también* Suplemento*, Tabla S1 y Figuras S1-S4). i) Clase baja: FFLs jerárquicos. Distinguimos aquellos FFLs en los que Z es genómicamente adyacente o no adyacente de Y. En la segunda de estas subfamilias se cuentan los casos explicables por homología. ii) Clase media: la menos poblada de las clases. Bajo el término* complejo *se agrupan los FFLs debidos a parejas (X, Y) en los que la acción de uno de los TFs depende totalmente de la presencia del otro miembro de la pareja. iii) Clase alta: FFLs formados en gran parte por la corregulación entre los hubs que hacen de X e Y. Distinguimos los casos estadísticamente significativos y, dentro de éstos, los atribuibles a homologías. El papel de CRP como elemento X es considerado a lo largo de toda la clasificación. B-C) Lógica reguladora dual implementada como FFL jerárquico (B) o policistrón autorregulado (C). El código de colores representa genes con funciones equivalentes.*

regulación adicional sobre esta misma zona por parte de otro TF puede afectar a los dos operones divergentes (Figura 3.3.E). Y si el único operón regulado por Y es el adyacente, este FFL constituido así por determinantes genómicos asignaría el máximo valor de FFLness ($\mathcal{F} = 1$) al TF que hace de Y.

Aunque el fenómeno de la *regulación del vecino* se diera (que no se da, como hemos visto) con la misma frecuencia sobre alguno de los operones regulados por un TF de alta conectividad, tendría unas consecuencias menos drásticas sobre la FFLness de éste, puesto que afectaría sólo a una fracción pequeña de su regulón. Por tanto, nos encontramos con un buen escenario neutral que de un plumazo podría explicar la alta FFLness de los TF de baja conectividad y su tendencia a la autorregulación.

De hecho, si el vínculo entre autorregulación y alta FFLness no era estadística-

mente concluyente, sí que lo es cuando comparamos la FFLness media de los TFs de baja conectividad que regulan algún operón vecino ($\mathcal{F} = 0.70$) con la de los que regulan sólo de un modo no adyacente ($\mathcal{F} = 0.29$, $p < 0.01$). Pero ante la constancia de este vínculo, ¿podríamos aducir aún que que la arquitectura genómica es consecuencia y no causa de la constitución del FFL con Y autorregulado? –esto es, ¿se han seleccionado estas arquitecturas divergentes porque dan lugar con facilidad a la constitución del motivo de red FFL, "verdadero" objeto de selección? En contra de esta interpretación estaría el hecho de que la regulación adyacente también se da entre los TFs del nivel superior de la red de regulación, es decir, no sujetos a regulación transcripcional adicional externa, donde la disposición divergente vuelve a ser la dominante y donde la regulación del vecino va ligada a la autorregulación de un modo significativo (p=0.01). Por tanto, la regulación adyacente es un fenómeno que en sí ya es objeto de selección, sea porque ello permite una localización más eficiente del sitio de unión al ADN por parte del TF [28] o porque da pie a que el fenómeno de transferencia horizontal incorpore a la célula funcionalidades autónomas (teoría del operón egoísta, *selfish operon* [90]). En este este segundo caso, parece que, tras la transferencia horizontal de grupos de genes en los que existe ya una regulación específica tipo $Y \rightarrow Z$, puede establecerse con posterioridad una regulación adicional externa [83], lo que podría dar lugar a la formación del FFL. La regulación externa adquirida es ejercida normalmente por un hub [83] (véase más abajo).

Como hemos dicho, en caso de duda, vamos a optar por el escenario neutral. Más aún, vamos a considerar explicados por la arquitectura genómica aquellos casos de regulación de un operón vecino en los que la disposición relativa de los dos operones no es divergente –es decir, que sea o convergente o unidireccional–, ante la posibilidad de que esta configuración provenga del cambio de orientación de una disposición divergente ancestral.

Aún así, la arquitectura genómica puede dar cuenta de la alta FFLness sólo de un modo parcial (34 de 74 casos, Figura 3.4.A) puesto que los TFs con regulación externa y con regulones pequeños muestran una fuerte tendencia a formar FFLs también con aquellos otros operones que no son regulados adyacentemente (Figura 3.5 y *Suplemento*, sección 4, Figuras S1-S2).

Segundo escenario neutral: ¿se explica la alta FFLness por fenómenos de homología?

Para los 40 casos restantes planteamos un segundo escenario neutral, en el cual los motivos podrían surgir de la duplicación de alguno de los elementos que los constituyen –de manera que con la duplicación se duplica también la conexión reguladora [91,92]. Se consideraron dos tipos de potenciales homologías: i) entre los TFs que hacen de X y de Y en cada FFL y ii) entre cada elemento Z no adyacente (nad$Z$) y alguno de los genes codificados en la *unidad central*, de manera que

la regulación a distancia fuera heredera de la establecida de manera local. Esta unidad central la definimos como el grupo de genes constituidos por el operón al que pertenece el TF que hace de Y y aquellos Z-operones localizados en el vecindario de aquél. Si se encontraban homologías de este tipo, el FFL involucrado era catalogado entre los neutrales. De nuevo, esta manera de proceder vuelve a ser muy conservadora (proclive a lo neutral) puesto que en i) estamos asumiendo que la duplicación sucedió después de que se estableciera la regulación y en ii) no se considera la influencia que en la conservación del enlace tiene la posición del gen duplicado en cada operón –es más plausible que con la duplicación del gen se herede la zona de regulación si el gen duplicado está codificado justo tras esta zona [83]. Finalmente, un total de 15 FFLs fueron de este modo descartados (Figura 3.4.A).

Tras la criba neutral

Tras esta doble y severa criba aún quedaba un número no desdeñable de FFLs[16] que se resistían a ser explicados bajo estos escenarios neutrales. Por ser los candidatos a los que con más fiabilidad se les podía asignar un origen adaptativo, nos planteamos si estos FFLs remanentes presentaban características funcionales comunes que permitiesen rastrear las razones de su selección. Y, efectivamente, estos FFLs presentan un patrón funcional muy claro ligado principalmente a la represión catabólica de los metabolismos alternativos al de la glucosa: en primer lugar, CRP actúa en la mayoría de los casos (20/25, Figura 3.4.A) como el elemento X del FFL; en segundo, la función de los genes codificados en los nad$Z$s está íntimamente relacionada con la de aquéllos de la correspondiente unidad central. Por ejemplo, nos encontramos que esta unidad y sus nad$Z$s pueden estar codificando transportadores alternativos de alta y baja afinidad por un mismo azúcar (Figura 3.5), distintos componentes de heteromultímeros o rutas metabólicas complementarias (véanse la sección 4 y el Apéndice del *Suplemento*).

La estrecha relación funcional entre los genes adyacentes y no adyacentes pone en tela de juicio la etiqueta de neutralidad adjudicada severamente a los FFLs que se forman con los primeros, puesto que unos y otros constituyen un conjunto unitario. La arquitectura transcripcional de este conjunto se presenta en la forma de agregado de FFLs que comparten sus elementos X e Y (Figura 3.5). Esto hace, naturalmente, que el predominio de CRP en el rol de X sea extensible a todo el conjunto de FFLs con Y de baja conectividad (40/74). Este predominio es estadísticamente significativo ($p = 0.008$) cuando se compara con el papel menos relevante de CRP en el resto de FFLs.

Este modelo unitario se ve además soportado por una doble vía: en primer lugar, con independencia de su localización respecto a Y, una dinámica común atribuible a la arquitectura FFL ha sido identificada para los elementos Z del

[16]En concreto, un tercio de los que tienen una Y de baja conectividad (25/74, Figura 3.4.A).

sistema de la arabinosa *araBAD* (adyacente) y *araFGH* (no-adyacente) [22], Figura 3.5.A. Por otro lado, nosotros hemos encontrado que la coconservación filogenética promedio de los pares (Y, Z) entre las γ-proteobacterias es significativamente mayor de la esperable a partir emparejamientos aleatorios. Además, cuando se estudian los casos adyacentes y no adyacentes por separado, la diferencia en esta coconservación no es significativa (*Suplemento*, sección 4).

La constitución de FFLs con los nad$Z$s rompe la imagen de FFLs surgidos como meros aspectos colaterales de la regulación adyacente. Probablemente, ésta se dé, como fenómeno previo al ensamblaje de los FFLs, por las razones que hemos tratado más arriba. Sin embargo, la evolución, en su *tinkering* continuo, no deja de aprovechar una circunstancia tan favorable a la constitución de lógicas duales (véase más abajo). A posteriori, esto supondría un beneficio adicional que posiblemente está contribuyendo a que la selección natural mantenga estas regulaciones locales.

3.4.2. El agregado de FFLs jerárquicos como unidad funcional adaptativa. Equivalencia con el policistrón autorregulado

El cuadro general que se nos presenta entre los FFLs con Y de baja conectividad es el de una serie de operones sujetos a un esquema de regulación jerárquica que combina la acción de dos TFs: uno general (X) y otro específico (Y) que se subordina al primero. Este esquema va ligado en la mayoría de los casos al fenómeno de la represión catabólica: cuando no hay glucosa la regulación por CRP activa una serie de genes cuyos productos son capaces de sentir (el Y del FFL) o metabolizar (los Zs, estén o no estén adyacentemente localizados) un azúcar alternativo [93]. Por tanto, agregados de FFLs como los de la Figura 3.5 se nos revelan como unidades funcionales independientes, en las que el cometido de cada elemento del motivo está claramente definido en el conjunto, y que han sido seleccionadas como vía de implementar la lógica jerárquica (X, Y). La autorregulación del elemento Y implica que la misma lógica de regulación dual aplica tanto al propio operón que codifica Y como a los Z-operones, pues todos los genes de Y y Z son regulados por la pareja (X, Y).

En esto último puede radicar la clave de la vinculación de la autorregulación con los Y de baja conectividad. Nótese que el FFL con Y autorregulado no es la única vía de implementar esta lógica en la que tanto el TF específico como las herramientas metabólicas son objeto de la regulación dual, pues existe una alternativa: el policistrón autorregulado con una regulación externa adicional (Figura 3.4.B-C). En este caso, tendríamos una serie de genes que, en lugar de actuar como Z en un FFL, serían parte del policistrón y estarían sujetos a la misma lógica jerárquica que hemos visto en los FFLs, sólo que ahora la señal específica vendría

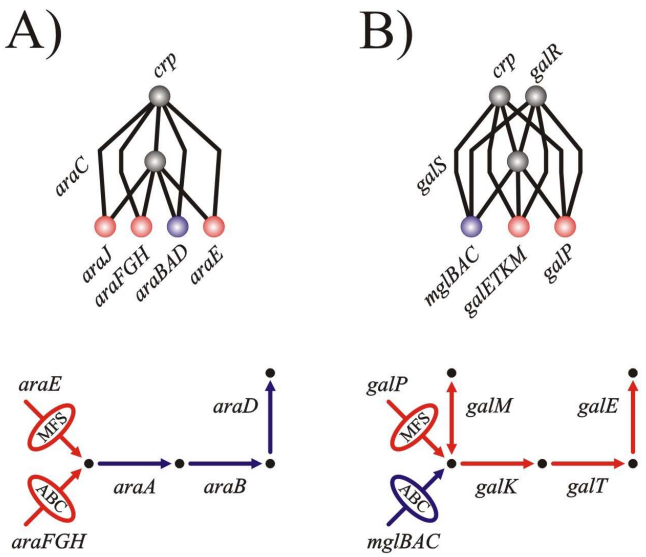


Figura 3.5: *Ejemplos de ajuste fino en los elementos Z de los FFLs jerárquicos: sistema de la arabinosa (A) y la galactosa (B). Arriba: En cada caso pintamos tanto las conexiones reguladoras entrantes y salientes del TF que se liga al azúcar específico como aquéllas involucradas en el ensamblaje de FFLs (conexiones X/Z). Nótese que ambos ejemplos exhiben un valor máximo de FFLness, $\mathcal{F} = 1$. Abajo: Rutas metabólicas codificadas. Flechas y elipses –cruzadas por flechas– denotan enzimas y transportadores, respectivamente. Se indica también el tipo de transportador (MFS o ABC). Código de colores para los elementos Z: azul, adyacente a Y; rojo, no adyacente a Y. El mismo código rige sobre las rutas metabólicas. Nótese como funciones equivalentes pueden estar codificadas tanto de manera adyacente como no adyacente. En el Apéndice del* Suplemento *figuran más ejemplos y se profundiza en la discusión de las funciones metabólicas codificadas.*

de la mano de la autorregulación del policistrón.

Que lo que se selecciona fundamentalmente es esta lógica por encima del modo en el que está implementada parece soportado por el hecho de que los policistrones autorregulados en los que, además, el TF específico codificado en el mismo no regula a ningún operón vecino, son mucho más frecuentes y codifican en promedio a un mayor número de genes cuando *existe* la regulación adicional externa[17].

La equivalencia entre los FFLs jerárquicos y los policistrones autorregulados también puede verse en ciertos resultados experimentales. En concreto, las medidas de expresión en uno de los elementos Z del sistema *gal* de la Figura 3.5.B muestran la misma aceleración de la respuesta que se atribuiría a un policistrón autorregulado negativamente [23]. De hecho, el elemento Y correspondiente está ne-

[17] Consúltese en el *Suplemento* las Tablas S3-S5 y la sección 4, donde se tratan las razones que pueden llevar a la selección de una u otra implementación.

gativamente autorregulado y, por tanto, comparte la misma lógica transcripcional del elemento Z.

Fijémonos ahora en el operón *lacZYA* (Figura 3.6), cuyo control transcripcional es un ejemplo paradigmático de estas regulaciones duales metabólicas. En este caso, sobre el gen que codifica el TF específico, *lacI*, no actúa regulación transcripcional alguna: ni la de CRP, ni la del propio represor LacI. Por tanto, ni la autorregulación ni la formación de FFLs son indispensables para el funcionamiento de la represión catabólica: lo único esencial es que la lógica dual actúe sobre los genes que codifican las herramientas metabólicas, sean enzimas (como LacZ) o transportadores (caso de LacY).

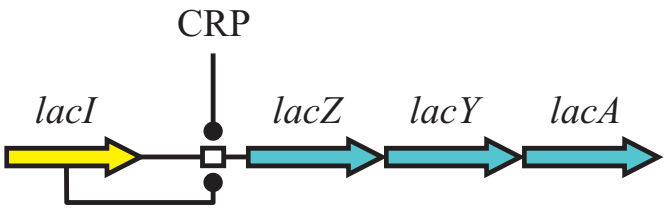


Figura 3.6: *Representación esquemática de la regulación dual sobre el operón lacZYA. El gen lacI, codificado adyacentemente a éste, no se autorregula ni sufre regulación por parte del TF global CRP.*

Pese a todo, en muchos casos puede resultar ventajoso que el TF específico esté sometido a esta lógica y que, en lugar de transcribirse de manera constitutiva como *lacI*, sólo se exprese, a semejanza de lo que sucede con las herramientas enzimáticas, cuando las condiciones son propicias. Por ello, todo esto, más que oponerse a que la implementación de la lógica (X, Y) en la forma concreta de FFL sea sujeto de selección adaptativa, parece establecer una jerarquía en la identificación del propio objeto de tal selección según el orden siguiente: i) lógica jerárquica sobre las herramientas metabólicas, ii) lógica jerárquica extendida el propio TF específico a través de la autorregulación, y iii) el FFL o el policistrón autorregulado como implementaciones alternativas de ii).

En cualquier caso, por cada Y con una alto valor de $\mathcal{F}$ tenemos un agregado de FFLs que comparten un mismo elemento Y (y con frecuencia un mismo regulador X). Los resultados anteriores, que se unen a los basados en recuentos estadísticos, muestran a estos agregados como unidades funcionales aisladas (como módulos).

3.5. Los FFLs más allá de la lógica jerárquica

En la sección anterior hemos visto que la fuerte señal de FFLness en los TFs de baja conectividad y la tendencia hacia la autorregulación de los elementos Y de estos FFLs se explica en última instancia por la selección de una lógica reguladora jerárquica, la cual está las más de las veces asociada a la represión catabólica. Pero, ¿explica esto toda la señal general de la identificación del FFL como motivo de red? Recuérdese que pese a que en la Figura 3.3.B-D se producía una drástica caída en la FFLness con la conectividad de los TFs, los valores se mantienen por encima de los neutrales de una manera estadísticamente significativa, lo cual indica que al menos una parte de los FFLs así constituidos podrían haber sido objeto de selección adaptativa.

3.5.1. FFL con Y de conectividad media

El grupo de los FFLs que cuentan con elementos Y de conectividad media sólo engloba a 35 de los 272 FFLs de la red (Figura 3.4.A). Los FFLs se presentan de nuevo formando agregados que comparten un mismo elemento Y –y en muchas ocasiones también el X (*Suplemento*, Figuras S3-S4). La FFLness media en este grupo está aún bastante por encima de los valores neutrales (Figura 3.3.B-D). En esta ocasión una parte de la señal proviene de tres casos de elementos Y que exhiben el valor máximo $\mathcal{F} = 1$ y que coinciden con respectivos pares (X, Y) en los que la acción de uno de los TFs depende totalmente de la presencia de su compañero[18]. Obsérvese que esta dependencia implica que el control autógeno, de existir sobre el elemento Y, ha de ir inevitablemente asociado a la constitución de los FFLs y que, en ese caso, la FFLness de Y no puede dejar de ser máxima (*Suplemento*, Figura S3).

3.5.2. FFL con Y de conectividad alta y bi-fan: las dos caras de la corregulación

Finalmente, tenemos el caso de los FFLs en los que el elemento Y es un hub. En este caso, la FFLness, aún permaneciendo significativa, sí que está cerca de los valores neutros (Figura 3.3.B-D). Sin embargo, esto no quiere decir que tales TFs no establezcan un número grande de FFLs en términos absolutos –de hecho, más de la mitad de los FFLs de la red tienen un hub como elemento Y–, sino que sólo se constituye una pequeña fracción de los FFLs potenciales. ¿Tiene entonces su origen esta familia de FFLs en la acción de fuerzas neutrales?

[18]El factor sigma RpoN depende de la presencia de NtrC para poder regular, RcsA depende de RcsB para formar un heterodímero; y CytR lo hace de CRP a la hora de formar un complejo regulatorio en la zona operadora. En conjunto, los tres casos dan lugar a la constitución de 19 de los 35 FFLs de esta categoría (Figura 3.4.A).

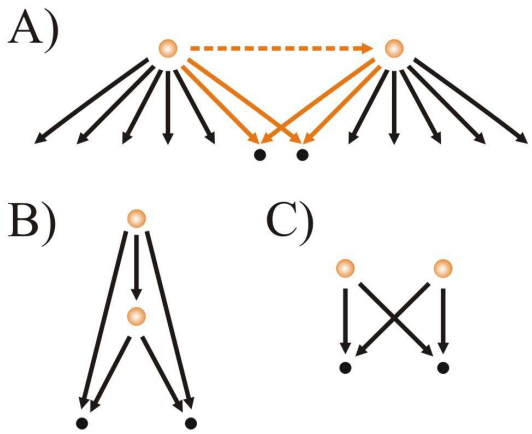


Figura 3.7: *Ensamblaje de motivos por solapamiento aleatorio de regulones grandes. A) En este ejemplo la regulación de un par de operones es compartida por dos hubs (nodos naranja). Si existe una conexión transcripcional entre los hubs esta corregulación da lugar a la formación de un agregado de FFLs (B). En caso contrario, se forma un bi-fan (C).*

A partir de la definición de FFLness tenemos que el número de FFLs que de una manera neutral puede ensamblar en promedio un TF actuando de Y viene dada por el producto de la FFLness neutral $\mathcal{F}_{neu}$ (de acuerdo al modelo nulo, $\mathcal{F}_{neu} \approx 0.08$, Figura 3.3.D), el número de regulaciones externas n_{in} y el tamaño del regulón n_{out} (Figura 3.3.A). Así, el ensamblaje nulo de FFLs escala con el tamaño del regulón y, por tanto, la baja FFLness de los Y con alta conectividad puede asociarse aún con la aparición de un número considerable de FFLs. Además, el papel de X suele ejercerlo casi siempre también un hub[19]. Parece plausible, pues, que una parte de los FFLs haya surgido a raíz del solapamiento aleatorio de los grandes regulones de los hubs (Figura 3.7). De todos modos, puesto que la FFLness de los hubs aún se mantiene en valores significativos, veremos que no se les puede atribuir este origen neutral a todos los FFLs constituidos por estos reguladores globales.

En seguida nos ocuparemos de esta cuestión. No obstante, antes de seguir, detengámonos en la Figura 3.7. En ella podemos ver que el solapamiento de los regulones da lugar a agregados de FFLs (que comparten los elementos X e Y) cuando uno de los hubs regula al otro. En cambio, si este vínculo no existe se constituye el último motivo de red que quedaba por tratar: el bi-fan y sus agregados (siempre que el solapamiento de los regulones implique a dos o más operones). Por tanto, el solapamiento aleatorio de los regulones de los hubs siempre contri-

[19]Esto ocurre en 211 de los 232 FFLs de la red; en particular, en 119 FFLs tanto el elemento X como el Y son hubs (Figura 3.4.A).

buirá al recuento de uno u otro de estos motivos de red; luego estudiando estos solapamientos en las redes aleatorias podemos lidiar a la vez con la significatividad estadística de aquellos FFLs y bi-fanes cuyos dos reguladores X e Y son hubs.

pares de TFs†		cre	cre_n	Z-score*	FFL / Bi-fan **
SoxS	MarA	10	66.7	17.68	→
FliA	FlhDC	5	40.8	10.52	←
FNR	NarL	18	42.4	9.97	→
FNR	ArcA	16	27.3	5.21	→
RpoE	CpxR	7	21.4	4.74	←
IHF	RpoN	10	22.6	4.58	n.i.
FNR	IHF	18	25.9	4.39	n.i.
IHF	NarL	8	23.3	4.28	n.i.
IHF	Lrp	7	19.9	3.63	n.i.
CRP	ArcA	19	38.2	3.56	n.i.
CRP	RpoE	2	3.6	-3.21	n.i.

Tabla 3.2: *Pares de hubs que exhiben una corregulación significativa.* † *En cada par, los TFs están ordenados por el tamaño de sus regulones respectivos.* * *El umbral de selección corresponde a* $p < 0.05$ *(corregido de comparación múltiple).* ** *Las flechas denotan el orden de regulación. cre: número de corregulaciones. n.i.: pares que no interactúan transcripcionalmente.* cre_n*: cre normalizada como* $cre/\sqrt{R_1 R_2}$*, siendo* R_i *es el tamaño del regulón. Nótese el caso de anticorregulación significativa entre CRP y RpoE (detalles en* Artículo*).*

Para investigar la relevancia de la corregulación neutral entre hubs identificamos, en primer lugar, todos los posibles emparejamientos de estos reguladores que, a priori, se pueden dar en la red de *E. coli*: 23 hubs, luego 253 parejas. Después, para cada una de las parejas contrastamos el número de corregulaciones observadas en *E. coli* con el valor nulo obtenido en las redes aleatorias (*Artículo*, Materiales y métodos, y *Suplemento*, sección 2). Finalmente, encontramos un pequeño número de corregulaciones significativas tras la corrección por comparación múltiple de los p-valores obtenidos [85] (Tabla 3.2 y *Suplemento*, sección 2). Nótese que los casos más significativos corresponden a cinco pares de hubs con interacción mutua y que, por ello, dan lugar a la formación de agregados de FFLs, mientras que los restantes pares de esta Tabla -que no interactúan- corresponden a agregados de bi-fanes [17, 18].

Si, tal y como procedimos con los FFL jerárquicos, adoptamos un escenario proclive de partida al modelo neutro tendremos en principio que descartar el par (MarA, SoxS) puesto que estos dos TFs son homólogos[20]. El conjunto de

[20]También consideramos si parte de estas corregulaciones significativas podrían haber surgido

corregulaciones no significativas ($p > 0.1$, incluso sin el ajuste por comparación múltiple) y aquéllas establecidas por el par (MarA, SoxS) implican un total de 73 de los 119 FFLs en los que tanto X como Y son hubs (Figura 3.4.A), lo que significa que, de acuerdo con el modelo neutro considerado, estos 73 FFLs podrían tener su origen en procesos neutrales. Esto concuerda con el valor cercano a lo neutral de la FFLness promedio de los elementos Y de alta conectividad. Las pocas parejas que establecen corregulaciones significativas serían las responsables de que este valor se mantenga aún por encima de los neutrales (Figura 3.3.B-D).

El grado de significatividad de la corregulación se refleja en el valor de $\mathcal{F}$ del elemento Y correspondiente. Efectivamente, la FFLness media de los Y que forman parte de corregulaciones que dan lugar a FFLs significativos (descontando a MarA, esto es, FliA, NarL, ArcA y RpoE) es $\mathcal{F} = 0.34$. En cambio, un valor promedio mucho más pequeño ($\mathcal{F} = 0.07$) se observa para el resto de hubs que tienen regulación externa y que no establecen corregulaciones significativas.

Finalmente, es interesante señalar que una de las parejas cuyas corregulaciones se encuentran entre las de más significatividad estadística es la de (FliA, FlhDC), que controla la expresión de los constituyentes de los flagelos de *E. coli* y su ensamblaje. El estudio de la dinámica de este sistema, considerado como agregado de FFLs, ha sido abordado experimentalmente [94]. Si a esto unimos el mismo tipo de aproximación realizado sobre FFLs jerárquicos [22–24], y nuestro propio tratamiento del sistema SOS [20], tenemos la circunstancia coincidente de que todos estos trabajos experimentales se han hecho sobre motivos naturales que entran dentro de los casos que hemos considerado más proclives a ser explicados bajo un escenario adaptativo. Posiblemente, esto ha contribuido a que en todos ellos se haya podido atribuir al motivo correspondiente el comportamiento de un ente funcional autónomo.

---

de la duplicación de los genes corregulados [78]. Este podría ser en parte el caso de las corregulaciones establecidas por IHF, FNR y NarL –que comparten 8 operones en sus respectivos regulones–, pero de nuevo sólo bajo un criterio permisivo puesto que, como ya hemos comentado, estamos aceptando la herencia de los sitios de regulación con independencia de la localización de los genes duplicados en sus respectivos regulones. Por otro lado, y por razones expuestas más arriba, no ha lugar a la consideración entre hubs de la arquitectura genómica como elemento condicionante de la constitución de los motivos.

Capítulo 4

La autorregulación en la búsqueda de un código de reconocimiento proteína/ADN en la familia de dominios hélice-giro-hélice/LacI

4.1. La posibilidad de un código parcial. Estrategias

La especificidad de la acción reguladora de un factor de transcripción (TF) descansa en el reconocimiento por su parte de cortas secuencias de nucleótidos; estas secuencias constituyen así los sitios de unión (*binding sites*, BS) del TF a la zona operadora del ADN que antecede a la que codifica cada operón regulado. La secuencia de nucleótidos por la que la proteína tiene afinidad no depende exclusivamente de los pocos aminoácidos de la misma que entran en contacto con el ADN a través de puentes de hidrógeno y enlaces van der Waals. Ello se debe a que la interacción entre aminoácidos y nucleótidos es un problema tridimensional fuertemente condicionado por la propia estructura estereoquímica de la proteína, que determina la geometría del contacto y, con ello, la energía libre del emparejamiento [39]. Esta es la razón de que todo intento de encontrar un código universal de reconocimiento proteína/ADN esté abocado al fracaso [38]. No obstante, la situación puede ser distinta en el caso de circunscribir el problema a un conjunto de proteínas cuyos aminoácidos de contacto con el ADN estén bajo un contexto estructural semejante. Entonces sí que se puede intentar la búsqueda de soluciones parciales a dicho código que sean válidas en el seno del conjunto en cuestión [39]. Nosotros hemos encontrado tales soluciones para un subconjunto importante de

la familia de TFs semejantes a LacI, cuyo dominio hélice-giro-hélice (*helix-turn-helix*, HTH[1]) es el principal responsable de la especificidad en el reconocimiento del BS adecuado.

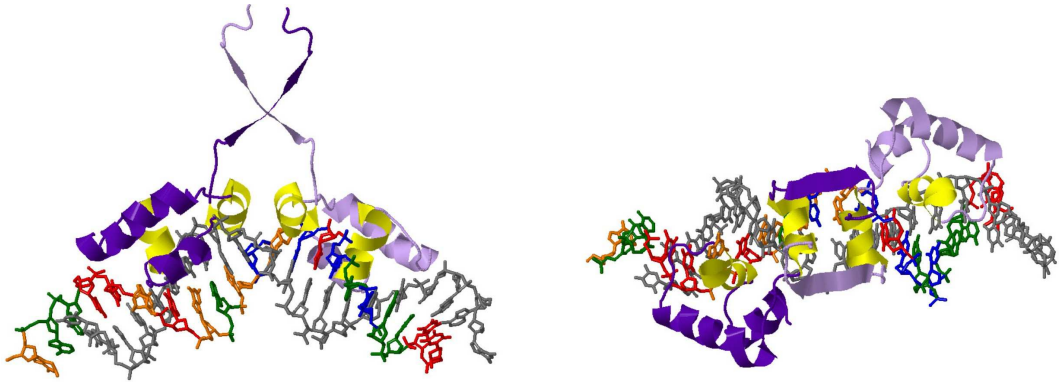


Figura 4.1: *Perspectivas frontal y superior del modelo estructural obtenido por difracción de rayos-X de la unión de un dímero de LacI al BS palindrómico SimL [95]:* `5'-GAATTGTGAGC·GCTCACAATT-3'` *(véanse también las Tablas 4.1 y 4.5). Una de las hebras de ADN está resaltada con el mismo código de colores que, para los nucleótidos, se usa en los logos de los alineamientos de BSs (A, verde; C, azul; G, naranja; T, rojo). Sólo representamos los aminoácidos que ocupan el tramo de posiciones 3 a 71 en la secuencia de LacI, lo cual abarca la práctica totalidad de las posiciones del alineamiento de dominios HTH-LacI de la Figura 4.3 (posiciones AA-1 a AA-69). En cada monómero hemos resaltado en amarillo la hélice bisagra (hacia el centro del BS), que enlaza con el grupo CpG central localizado en la ranura menor del ADN. Con el mismo color aparece la hélice de reconocimiento (más hacia los extremos del BS) que realiza contactos de naturaleza específica con los nucleótidos de la ranura mayor. Figura realizada con Jmol a partir de la estructura con PDB ID: 1lbg [95].*

Por su importancia histórica en el desarrollo del modelo del operón, la proteína LacI ha sido uno de los pocos miembros de la familia HTH en los que se ha resuelto por cristalografía de rayos-X o resonancia magnética nuclear (RMN) el complejo formado por la unión del dominio al ADN [37]. El modelo derivado de estos estudios [30–34] y soportado por experimentos con variantes mutados de LacI [35, 36] hace recaer gran parte de la especificidad en tres de los aminoácidos de la hélice de reconocimiento (HR), que es la segunda hélice del dominio HTH, leído desde el extremo N-terminal (Figura 4.1).

El ritmo exponencial con el que los genomas bacterianos son secuenciados ha hecho crecer hasta varios millares el número de proteínas en las que se ha anotado un dominio HTH-LacI. Aunque con ello ya tendríamos la parte del código correspondiente a los aminoácidos, en muchos casos los BSs no están identificados. Esta es, por tanto, una cuestión previa crucial que se ha de solucionar y que

[1]Véase la sección de *Materiales y métodos* en este mismo Capítulo.

constituye un problema en sí mismo. Y dada la ingente cantidad de genomas secuenciados, la identificación a gran escala de los BSs (o de sitios candidatos a serlo) ha de descansar obligatoriamente en métodos bioinformáticos, de entre los que sobresale el de las huellas digitales filogenéticas (*phylogenetic footprinting*, PF) [96]. Este método se basa en el hecho de que los BSs, a causa de su funcionalidad dependiente de secuencia, sufren una presión selectiva mayor, por lo cual están mejor conservados que el resto de la zona no codificante.

El punto clave del PF es la selección de las secuencias que se van a comparar y sobre las que se va a realizar la búsqueda de los BSs. El método habitual, que consiste en tomar las regiones no codificantes que anteceden a genes ortólogos, asume implícitamente que genes ortólogos son regulados por TFs ortólogos, los cuales conservarían la afinidad por un mismo tipo de secuencia de nucleótidos. Esta aproximación, que pierde eficacia conforme crece la distancia evolutiva, ha de enfrentarse a una identificación de los ortólogos muchas veces problemática debido a los fenómenos de duplicación y de pérdida de genes [29]. Además, ya la propia definición de ortología no implica necesariamente la conservación de la funcionalidad [29]. Por otro lado, y en relación con la cuestión que nos ocupa, basta que dos TFs ortólogos discrepen precisamente en alguno de los aminoácidos críticos de la HR para que su afinidad por una secuencia dada difiera sensiblemente.

En el capítulo anterior, una de las circunstancias barajadas como posible condicionante del ensamblaje de los motivos de red fue la arquitectura genómica –en concreto, la posible localización adyacente del operón que codifica a un TF[2] y de alguno de los operones regulados por el mismo. Este tipo de regulación adyacente se da mucho más entre los TFs específicos [21, 97] –esto es, los que regulan a muy pocos operones– como sucede en el caso de aquéllos involucrados en metabolismos alternativos al de la glucosa. También vimos que la regulación adyacente está muchas veces asociada a la propia autorregulación a través de la disposición divergente de los operones regulador y regulado, aunque también es frecuente la regulación adyacente en la que el operón regulado está codificado corriente abajo en la misma hebra que el regulador [97] (como en el paradigmático caso del operón que codifica LacI y el operón regulado *lacZYA*).

La profusión de la autorregulación en la red de transcripción de *E. coli* -el 56 % de los TFs está autorregulado [21]- y, en general, de la regulación adyacente entre los TFs específicos procariotas, animaba a proponer una estrategia alternativa para la identificación de los BSs que, en lugar de basarse en las relaciones de ortología, lo hiciera sobre estos fenómenos de regulación local, buscando los BSs en las regiones no codificantes que anteceden al operón del TF y al operón codificado inmediatamente a continuación[3] (Figura 4.2).

[2]Operón al que para abreviar designaremos como *operón regulador*.

[3]Siempre y cuando éste se encuentre en la misma hebra, lo cual sucede en promedio en la mitad de los casos. No consideramos, por tanto, la disposición alternativa convergente, en la que el operón situado corriente abajo está codificado en la hebra contraria. Bajo esta arquitectura

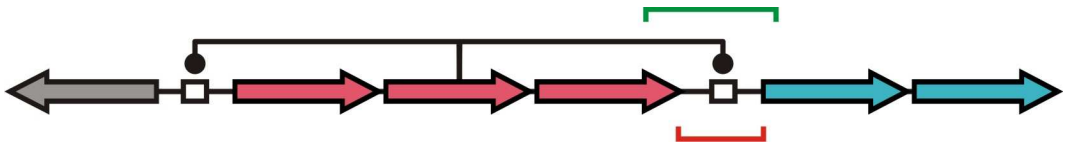


Figura 4.2: *Búsqueda de BSs basada en la regulación adyacente. Los colores de los genes indican aquéllos pertenecientes a un mismo operón. El gen que codifica un TF (el del centro del operón regulador, en rojo) autorregula la transcripción de su propio operón. Esta autorregulación puede ir unida a la regulación de un operón situado corriente arriba en disposición divergente (en gris). Cuando el TF regula a un operón situado corriente abajo, éste suele estar codificado en la misma hebra que codifica a aquél (disposición unidireccional). La línea roja marca una región de búsqueda de los BSs (cajas blancas) que es estrictamente intergénica. La linea verde es una versión extendida de la anterior* (Materiales y métodos).

Por cada TF de la familia HTH-LacI dispondríamos así de una o dos secuencias intergénicas para la búsqueda de BSs; secuencias que, de cara a la aplicación de la técnica del PF, hemos de agrupar en subconjuntos que potencialmente puedan albergar un mismo tipo de BS. La técnica habitual de agrupamiento de ortólogos parece apostar por contextos estructurales semejantes con independencia de las diferencias que puedan existir en los aminoácidos que contactan la doble hélice. Nosotros vamos a realizar una apuesta distinta, agrupando todos los TFs que, dentro de la familia HTH-LacI, comparten una misma secuencia en las posiciones clave de la hélice de reconocimiento.

A diferencia de los métodos basados en la ortología, el nuestro es un agrupamiento de definición estricta y que es independiente de la distancia filogenética entre las secuencias de las proteínas involucradas. Además, esta forma de agrupar las secuencias revelará la existencia del código buscado si el alineamiento de los BS que se encuentren tiene un alto contenido de información –como cabe esperar si es correcta la hipótesis de que el contexto estructural a lo largo y ancho de la familia va a ser lo suficientemente estable como para que los BSs queden determinados en gran medida por los aminoácidos críticos para la especificidad. Como se verá, esto es lo que sucede en la mayoría de los casos, por lo que hemos encontrado un código de referencia de gran cobertura dentro de esta familia y que, en consecuencia, rige sobre un subconjunto importante de los TFs procariotas. Las situaciones en las que el contexto estructural puede variar lo suficiente como para dar lugar a BS alternativos –y por tanto, emborronar el código– se nos muestran como las excepciones que confirman la regla.

la regulación adyacente se da con menos frecuencia [97].

4.2. Resultados

4.2.1. Secuencias de aminoácidos predominantes en la HR

Es evidente que la universalidad de cualquier código será tanto mayor cuantos más sean los taxones y TFs que se tengan en cuenta a la hora de buscar tal código. Por ello nos remitimos a la base de datos MicrobesOnline [98] –que en el momento de escribir este documento abarca un millar de genomas procariotas– para obtener todas las secuencias de dominios anotados HTH-LacI (Smart SM00354). Este dominio incluye tanto el motivo HTH propiamente dicho como la hélice-bisagra (Figura 4.1 y *Materiales y métodos*). La búsqueda de BSs por la técnica del PF exigía una cierta distancia filogenética entre los organismos a comparar. Además, existe una sobrerrepresentación de los genomas muy cercanos a los de los organismos modelos –en muchas ocasiones se trata de meras variaciones de estirpe. Por ello, se descartaron del conjunto inicial las secuencias redundantes para quedarnos finalmente con un total de 2639 dominios (*Materiales y métodos*). La Figura 4.3 muestra el logo resultante del alineamiento del conjunto completo de estos 2639 dominios HTH-LacI.

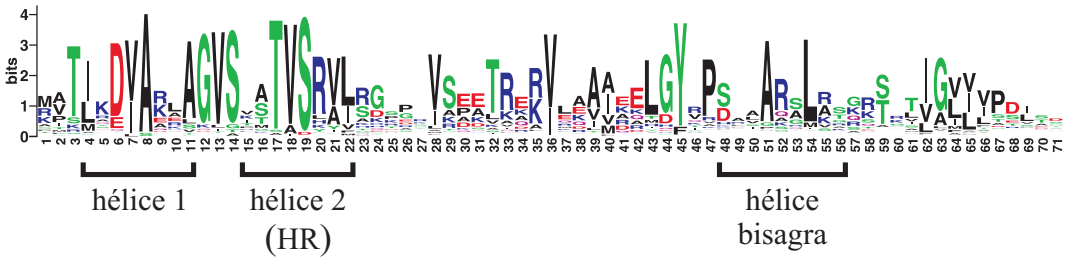


Figura 4.3: *Logo resultante del alineamiento de 2639 dominios HTH-LacI. Seguiremos el criterio de referirnos a aminoácidos de proteínas particulares por la posición (antecedida del prefijo AA) que dichos aminoácidos ocupen en este alineamiento. La hélice-1, la hélice-2 (o HR) y los residuos intermedios constituyen el motivo HTH propiamente dicho (véanse los* Materiales y métodos).

Por otro lado, en la Tabla 4.1 se cotejan los resultados de diversos trabajos de cristalografía de rayos-X y RMN que estudian el complejo formado por la unión del dominio (incluyendo la región de la hélice-bisagra) al ADN para varios miembros de la familia de TFs que nos ocupa (en concreto, LacI [32, 34] y PurR [30, 31] de *Escherichia coli*, y CcpA [33] de *Bacillus megaterium*). En esta Tabla se enumeran las semisecuencias para las que existe constancia de la interacción entre un aminoácido y un nucleótido dados. La semisecuencia correspondiente al lado derecho del BS natural O1 de LacI tiene un nucleótido más que el resto (véase también la Tabla 4.5). En este caso, uno de los dominios HTH-LacI de los que consta el dímero de LacI se desplaza alejándose una posición respecto al

		NT-8	NT-7	NT-6	NT-5	NT-4	NT-3	NT-2bis	NT-2	NT-1
1) LacI O1 izq.		T	T	G	T	G	A	-	G	C
2) LacI O1 der.		T	T	G	T	T	A	T	C	C
3) LacI SymL		T	T	G	T	G	A	-	G	C
4) LacI-$V_{15}A_{16}$ SymL*		T	T	G	T	A	A	-	G	C
5) CcpA izq.		C	T	G	T	T	A	-	G	C
6) CcpA der.		C	T	G	A	A	A	-	G	C
7) PurR *purF*-SymR		A	C	G	C	A	A	-	A	C
hélice-1	AA-4						1,3,4,7		1,3,4,7	
	AA-5								1,3,4	1
giro	AA-14			4	1,2,3,4,5					
HR	AA-15				4	1,3	1,2,3,4,5,6,7	2		
	AA-16			4	1,2,3,4,5,7	1,2,3,4,6,7	2			
	AA-19					4,7	2,3,4			
	AA-20		1	1,2,3,5,6	7					
	AA-26			7						
	AA-27	1,2,3,4	1,2,3,4,7							
hélice	AA-51									1,2,4
bisagra	AA-54									1,2,3,4,5,6,7
	AA-55								1,5,6,7	2,4,5,6

Tabla 4.1: *Interacciones de aminoácidos del dominio HTH-LacI con los nucleótidos (o sus complementarios en la hebra opuesta) de siete semisecuencias de BSs según se documenta en los siguientes trabajos de RMN y/o cristalografía: semisecuencias 1 y 2 en [32]; 3 y 4 en [34]; 5 y 6 en [33]; y 7 en [30] y [31]. La enumeración de las interacciones constatadas se hace sobre el número de semisecuencia. No se consideran las interacciones con los grupos fosfatos y desoxirribosas, que son de naturaleza no específica. Las coordenadas de aminoácidos y nucleótidos se anteceden de los prefijos AA y NT, respectivamente. La numeración de los nucleótidos empieza por los más cercanos al centro del BS. La conservada citosina en la posición NT-1 es la C del grupo CpG central, véase por ejemplo la Figura 4.7. Se han resaltado en rojo los nucleótidos presentes en todas las semisecuencias y también los aminoácidos (o grupos de aminoácidos semejantes) muy conservados (Figura 4.3).*

centro del BS de manera que, pese al intercalamiento de NT-2bis, realiza contactos con la semisecuencia derecha similares a los que realiza el otro monómero sobre la semisecuencia izquierda [32]. El interés por evidenciar la similitud de los contactos subyace al modo en el que se han alineado las semisecuencias en la Tabla 4.1.

Como se ve, en el caso de BSs no palindrómicos mostramos los contactos con las semisecuencias izquierda y derecha –estas últimas leídas en la hebra complementaria. Las semisecuencias provenientes de BSs palindrómicos son: SymL, simetrización de la semisecuencia izquierda del sitio de unión O1 de LacI de *E. coli* [32]; SymL*, variante de la anterior que la asemeja al BS natural de GalR/GalS [36]; y *purF*-SymR, simetrización de la semisecuencia derecha del BS natural desde el que PurR regula la transcripción de *purF*. LacI-$V_{15}A_{16}$ es un mutante de LacI con la misma secuencia de aminoácidos de reconocimiento [AA-15, AA-16] que GalR/GalS y que se une eficientemente a SymL* [34, 36].

De la comparación entre la Figura 4.3 y la Tabla 4.1 quedémonos en primer

lugar con el alto grado de conservación del residuo de leucina en la posición AA-54 del alineamiento de dominios –aparece en 2164 de las 2639 secuencias. Este residuo de la hélice-bisagra se intercala entre el grupo CpG característico del centro de sus BSs respectivos [37] (véase la Figura 4.1). La interacción de AA-54 con el nucleótido de la posición NT-1 de las semisecuencias (que corresponde a la C del grupo CpG) aparece documentada en todos los estudios recogidos en la Tabla 4.1. Otros residuos hidrofóbicos como la metionina (149 casos) pueden jugar el mismo papel de L_{54} [31, 99]. De hecho, los residuos que acaparan esta posición son todos hidrofóbicos –L, M, I, V, F, A (en orden de frecuencia)– lo que abarca al 93 % de los dominios. La muy conservada A_{51} está también relacionada con la unión de la hélice-bisagra al grupo CpG [100], sea sólo a través de su enlace no específico con grupos fosfatos [33], o mediante su interacción directa con los nucleótidos (Tabla 4.1). Este patrón de conservación de residuos clave de la hélice-bisagra entra en concordancia con la abrumadora presencia de tales grupos CpG en la relación de BSs que, para la familia HTH-LacI, han sido dilucidados bien experimentalmente o por procedimientos puramente informáticos, como recogen bases de datos de BSs tan completas como la RegTransBase (RTB) [101].

Si la hélice-bisagra se une a una zona CpG central, las diferencias en la especificidad por un tipo de secuencia u otra recae sobre todo en los aminoácidos de la HR, en concreto, se consideran residuos clave para la especificidad aquéllos que ocupan las posiciones AA-15, AA-16 y AA-20 [37]. La acción de este último sobre la posición NT-6 de la semisecuencia parece un fenómeno de especificidad independiente, como atestiguan los datos mutacionales [36] y las interacciones documentadas (Tabla 4.1). A diferencia de los otros dos casos, AA-20 muestra un notable grado de conservación; como consecuencia, el predominio de la arginina en esta posición se corresponde con el de la guanina en NT-6 del BS, con la cual establece un puente de hidrógeno [100].

Las secuencias consenso que se han venido elaborando para miembros de la familia HTH-LacI en distintos grupos de organismos apuntan la alta variabilidad de las posiciones NT-4 y NT-5 [102], variabilidad de la cual las semisecuencias que aparecen en la Tabla 4.1 sólo ofrecen una pequeña muestra. Gran parte de la especificidad del BS parece residir, pues, en estas posiciones. Confrontando esta misma Tabla con la Figura 4.3 podemos apreciar cómo, de los aminoácidos que interaccionan con NT-4 y NT-5, son precisamente los que ocupan las posiciones AA-15 y AA-16 los únicos cuya baja conservación a lo largo de la familia de dominios entra en consonancia con la variabilidad de estos nucleótidos.

El vínculo del AA-16 con NT-4 y NT-5 está unánimemente documentado en la Tabla 4.1. Sin embargo, esa unanimidad no se da en el caso de la interacción de AA-15 con las mismas posiciones. Además, llama la atención que, en cambio, en todas las semisecuencias este aminoácido se asocia con NT-3, que es una posición ocupada por una adenina de manera reiterada –es una de las tres posiciones que permanecen invariantes en las siete semisecuencias. Más esperable es la relación

de esta omnipresente adenina con posiciones de aminoácidos también muy conservadas: AA-4 y AA-19 (este último también en la HR). A tenor de todo esto, el residuo AA-16 parece tener una mayor relevancia como determinante de la especificidad, como también sugieren los estudios mutacionales [36]. En cualquier caso, es importante señalar que la sustitución de Y_{15} y Q_{16} de LacI de *Escherichia coli* por los residuos que ocupan las mismas posiciones en otros TFs de la familia HTH-LacI es capaz de recuperar la distinta especificidad de estos últimos, como en el mencionado mutante LacI-$V_{15}A_{16}$ [34, 36].

Entre los aminoácidos de la HR no asociados a especificidad se encuentra un grupo de tres (AA-17, AA-18 y AA-19) que sufren muy poca variación a lo largo de toda la familia de dominios (Figura 4.3) pues casi en el 80 % de los casos (2080 dominios) encontramos la secuencia TVS en estas posiciones (Figura 4.4.A). Si unimos esto al mencionado predominio de la arginina en AA-20, al final resulta que para el tramo de residuos de la HR que va de AA-17 a AA-20 la combinación TVSR está presente en 1490 de los 2639 dominios (56.5 %, Figura 4.4.B). Dicho con otras palabras: para una mayoría del conjunto de los dominios HTH-LacI no redundantes considerados en este trabajo el potencial de especificidad asociado a variaciones de la posición AA-20 no se utiliza. Por tanto, parece ser que, a priori, la discriminación entre el conjunto de BSs diferentes que podemos asociar a esta familia de dominios va a recaer en gran medida en los residuos AA-15 y AA-16.

En este trabajo nos hemos ceñido a este grupo mayoritario de dominios que presenta la secuencia TVSR para encontrar el hipotético código que, al menos en este importante subconjunto de la familia HTH-LacI, relaciona las posiciones AA-15 y AA-16 de los dominios HTH-LacI con las posiciones NT-4 y NT-5 de los BSs. La amplitud de este subconjunto supone que tal código se establece como una referencia de partida para los casos menos frecuentes en los que se dan secuencias alternativas a la TVSR.

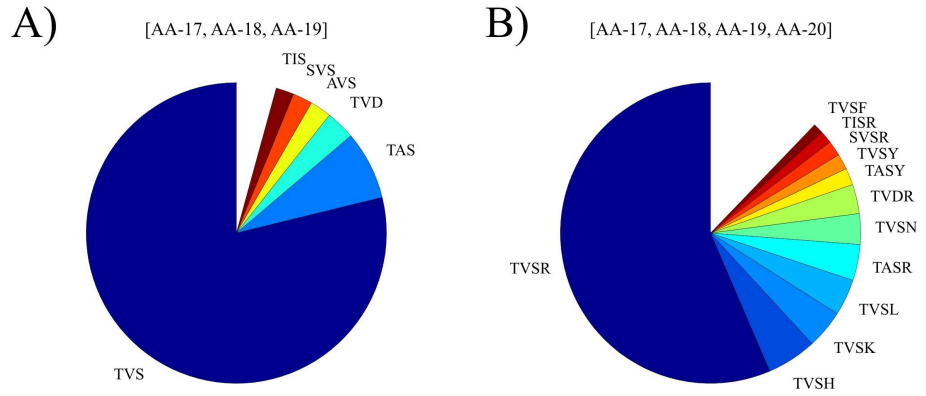


Figura 4.4: *A) Distribución de secuencias para la tríada de aminoácidos de la hélice de reconocimiento* [AA-17, AA-18, AA-19]. *Sólo constan las secuencias presentes en más de un 1 % de los dominios. B) Ídem para* [AA-17, AA-18, AA-19, AA-20].

4.2.2. Eficacia del método de búsqueda de BSs basado en la regulación adyacente

Como explicamos más arriba, las regiones intergénicas fueron agrupadas en conjuntos correspondientes a zonas adyacentes a dominios con una misma secuencia en las posiciones AA-15 y AA-16. Dentro de cada conjunto se realizó una búsqueda de BSs dividida en dos etapas: la primera, por PF sobre la versión estrictamente intergénica de las regiones de búsqueda (Figura 4.2); y la segunda, mediante la construcción y aplicación de la matriz de peso de las posiciones (*position weight matrix*, PWM) sobre la versión extendida (*Materiales y métodos*).

Las secuencias consenso parciales que se han venido construyendo para los BSs correspondientes a distintos subconjuntos de la familia HTH-LacI tienen un tamaño típico de 14 nucleótidos [102], pues para posiciones más alejadas del centro del BS que NT-7 la conservación de los nucleótidos es, en general, escasa. Por otro lado, ya nos hemos referido a la reiterada documentación del grupo CpG central en los BSs que se vienen resolviendo para la familia HTH-LacI, lo cual entra además en concordancia con la alta conservación de los residuos AA-51 y AA-54 de la hélice bisagra que lo contactan –aunque existen casos documentados en los que la ausencia del residuo hidrofóbico en AA-54 se corresponde con BSs alternativos carentes del CpG central [102], lo anterior apunta a que dichas alternativas tienen un carácter minoritario en el seno de esta familia.

Por tanto, todo parece indicar que el modelo estándar de funcionamiento de los dominios HTH-LacI es similar a los casos estructuralmente resueltos en los que la unión de un TF dimérico al ADN recae en i) la HR (depositaria en gran medida de la especificidad de la acción del TF) y que se une al surco mayor del ADN, y ii) la unión de la hélice-bisagra a un grupo CpG central, localizado en el surco menor (Figura 4.1). Sin embargo, para evitar problemas de circularidad en la búsqueda de los BSs (es decir, encontrar secuencias sesgadas hacia las ya conocidas) se hizo un uso bastante parco de toda esta información. Así, a la hora de fijar los parámetros del PF, todo lo anterior se tradujo en una búsqueda de BSs palindrómicos con un rango para las posiciones más conservadas en torno a los 14 nucleótidos (*Materiales y métodos*). No se utilizó información previa sobre secuencias de BSs ya documentadas experimental o informáticamente. Finalmente, el alto porcentaje de TFs para los cuales se han podido encontrar BSs –considerando que sólo se busca en las regiones aledañas a la del operón regulador– y la tipología de los mismos (ver más abajo) concuerdan con la hipótesis de que el modelo estándar mencionado rige sobre gran parte de los dominios HTH-LacI.

La principales ventajas del método de búsqueda de BSs basado en la regulación adyacente estriban en i) la posibilidad de su inmediata aplicación a todo genoma recién anotado –basta que la anotación incluya el dominio en cuestión y una predicción de los genes que se han de considerar dentro de cada operón–, y ii) la evitación de los problemas asociados a las relaciones de ortología y de

funcionalidad. Por ejemplo, resulta trivial que en el caso de la autorregulación existe relación funcional entre el TF y el operón regulado (véase el Capítulo de *Conclusiones*).

A priori, la principal desventaja de este método frente a los tradicionales consiste en la exclusión de la búsqueda de aquellos BS localizados en otras regiones del genoma diferentes al entorno de la que codifica el TF. Por tanto, en promedio, el número de BSs asociado a cada TF será menor en nuestra estrategia. Sin embargo, al poder efectuar la búsqueda de una manera automática (y sin las cortapisas mencionadas) sobre cada genoma secuenciado y anotado, el número total de BSs encontrados es superior al que se halla por la estrategia ordinaria, lo cual tiene mucha importancia para nuestro objetivo de encontrar el código proteína/ADN. Así, el número de TFs para los que encontramos al menos un BS (712 de los 1490 considerados) casi triplica a los 271 miembros de la familia HTH-LacI con la secuencia TVSR que figuran en la RTB (Figura 4.5.A), una vez se eliminan las redundancias de esta base de datos aplicando el mismo criterio usado con nuestras secuencias (*Materiales y métodos*). Nótese que estos 712 casos (que involucran 572 autorregulaciones y 207 regulaciones corriente abajo) cubren casi la mitad de los 1490 dominios considerados en este trabajo. Y efectivamente, el alto número de TFs considerados es capaz de compensar la limitación local de la búsqueda, puesto que el número total de BSs encontrados es también mayor (942 frente a 721 en la RTB, Figura 4.5.B).

4.2.3. Características generales de los BSs. Posiciones (casi) invariantes. Posiciones específicas

A partir del alineamiento de los BSs que encontramos asociados a una misma secuencia [AA-15, AA-16] se construyó el correspondiente logo. En la Figura 4.6 mostramos los cuatro casos que involucran a un mayor número de TFs. Antes de entrar en la deducción de un código de reconocimiento, analicemos las características generales del conjunto completo de los logos, el cual se puede encontrar en el Apéndice D, calculando la secuencia consenso del alineamiento subyacente a cada uno. Si se alinean a su vez estas secuencias consenso se obtiene el *logo-consenso* que aparece en la Figura 4.7 y que da cuenta, finalmente, de las características generales de los BSs para la familia HTH-LacI –recordemos que estamos considerando el subconjunto mayoritario que comparte la secuencia TVSR en la HR. La utilización de un alineamiento de secuencias consenso en lugar del alineamiento bruto de todos los BSs encontrados evita la sobrerrepresentación de los casos correspondientes a secuencias [AA-15, AA-16] que involucran a un mayor número de TFs.

El logo-consenso demuestra que el modelo de interacción revelado por los estudios con LacI, PurR y CcpA tiene vigencia como poco en una fracción importante de la familia HTH-LacI. Por un lado, el destacado grupo CpG central señala que

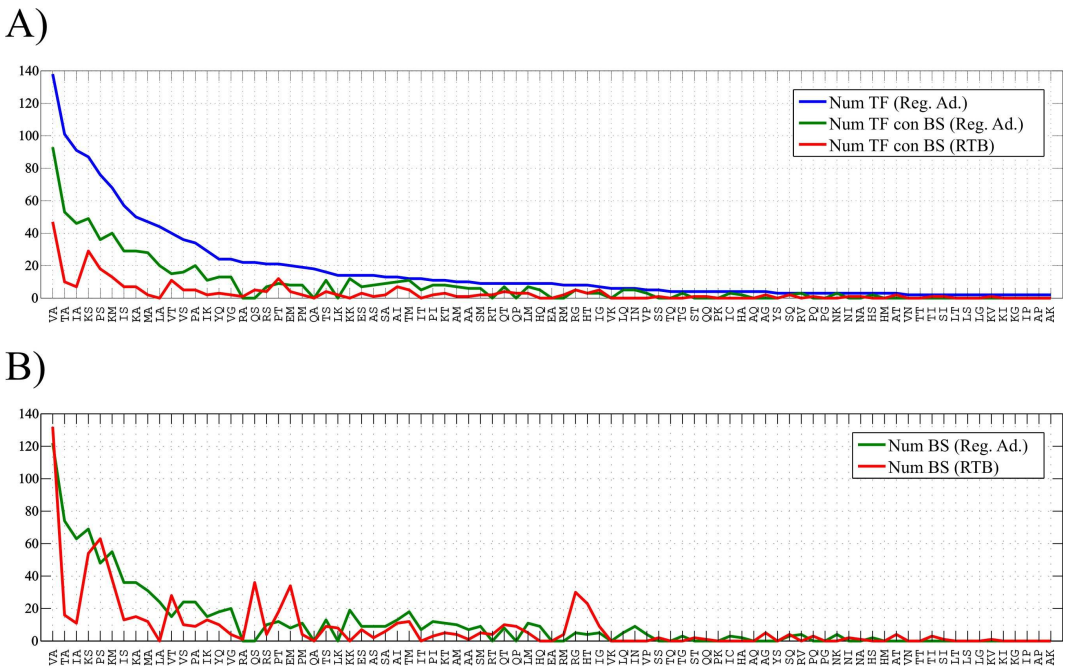


Figura 4.5: *Comparación con la RTB. Los datos se desglosan por dominios tipo TVSR con una misma secuencia en las posiciones* [AA-15, AA-16]*. A) línea azul: número de TFs presentes en nuestro conjunto de datos; línea verde: número de éstos para los que se encontró al menos un BS; línea roja: número de TFs presentes en la RTB –en esta base de datos los TFs siempre están asociados a BSs (Materiales y métodos). B) número total de BSs (línea verde: este trabajo, linea roja: RTB).* Reg. ad.*: regulación adyacente.*

el modo de acción de la hélice-bisagra sobre el BS se mantiene a lo largo la filogenia del mismo modo que se conservan sus residuos clave AA-51 y AA-54. Por otro, dado que todos los dominios que hemos considerado tienen una arginina en AA-20, la conservación de la guanina en NT-6_L (y la correspondiente citosina en NT-6_R)[4], que apenas se ve afectada de la gran variabilidad de AA-15 y AA-16, corrobora el carácter independiente del enlace específico de AA-20 sobre NT-6.

Resulta llamativa la fuerte conservación de NT-3. Esto ya fue señalado en la Tabla 4.1, según la cual este sitio es contactado por aminoácidos con muy distinto grado de conservación, destacando en todos los estudios estructurales referidos la documentación de contactos desde el muy variable residuo AA-15, lo que apunta a que el par adenina/timina en NT-3 viene determinado sobre todo por necesidades estructurales en el modo de actuar de este tipo de TF sobre sus BSs. Algo parecido sucede con NT-7 –aunque aquí la conservación de los nucleótidos

[4]Dado que en la mayoría de los casos los BS no van a ser palíndromos perfectos, es necesario considerar ambas semisecuencias; en ese caso, los subíndices L y R corresponden a coordenadas de las semisecuencias izquierda y derecha, respectivamente.

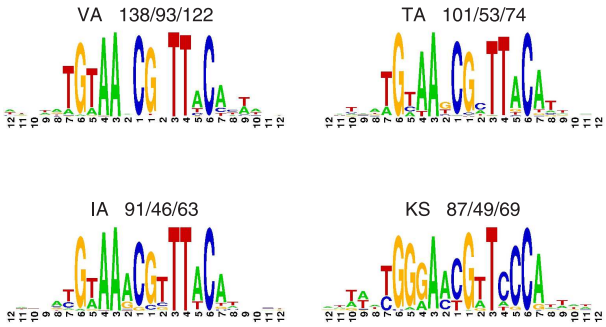


Figura 4.6: *Logos correspondientes a los BSs asociados a una misma secuencia de aminoácidos de reconocimiento* [AA-15, AA-16]. *Se muestran los cuatro casos en los que están implicados un mayor número de TFs (véanse las abscisas de la Figura 4.5). El resto de casos se puede consultar en el Apéndice D. Sobre cada logo se muestran la secuencia de los aminoácidos de reconocimiento y una tríada de números (i/ii/iii) correspondientes a i) el número total de TFs que tienen esta secuencia, ii) el número de TFs para los que se encontró al menos un BS y iii) el total del BSs encontrados. En las representaciones de logos omitimos el prefijo NT de la numeración de las posiciones.*

no es tan destacada– puesto que la Tabla 4.1 asocia el par timina/adenina en NT-7 con la posición AA-27, poco conservada en el alineamiento de los dominios (Figura 4.3).

Frente a las posiciones con un (cuanto menos) notable grado de conservación, la variabilidad de NT-4 y NT-5 corresponde a su condición de ser los determinantes por antonomasia de la especificidad y está correlacionada con la de los residuos que presumiblemente los contactan, AA-15 y AA-16. La clara delimitación de esta zona de no conservación en el seno del logo-consenso es la base de la existencia de un código de reconocimiento entre estas parejas de nucleótidos y aminoácidos.

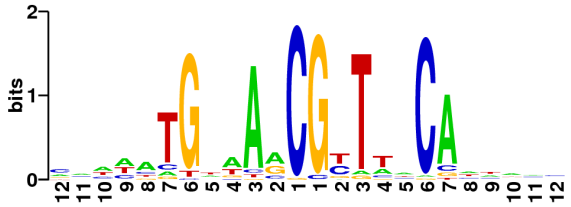


Figura 4.7: *Logo-consenso para los BSs asociados al conjunto de miembros de la familia HTH-LacI con secuencia TVSR en la HR. En el caso de los BSs correspondientes a TFs con los mismos aminoácidos de reconocimiento que LacI de E. coli ($Y_{15}Q_{16}$), no se ha considerado la posición insertada NT-2bis (véanse la Tabla 4.1 y el Apéndice D).*

4.2.4. Un código dos a dos. Ambigüedades. Degeneración en AA-15

Cuando en un BS existen posiciones que no están ocupadas precisamente por aquellos nucleótidos por los que el TF tiene una especial afinidad, la probabilidad del enlace del TF al BS disminuye [103]. Esta disminución no tiene por qué conllevar siempre la inoperatividad del BS, puesto que la suma de afinidades por el resto de posiciones del BS aún podrían ser compatibles con la acción reguladora, si bien ésta tendría un carácter más laxo. Sin embargo, un código de reconocimiento proteína/ADN sólo ha de recoger aquellas secuencias de nucleótidos NT-4 y NT-5 que realmente impliquen especificidad en la acción reguladora y que, por tanto, estén sujetas a procesos de selección purificante (*purifiying selection*). Por ello, sólo fueron consideradas como secuencias candidatas a formar parte del código aquéllas cuya presencia tenía un mínimo de significatividad estadística frente a un modelo nulo en el que las secuencias surgen como combinaciones neutrales de los cuatro nucleótidos (*Materiales y métodos*).

Sin embargo, con mucha frecuencia las combinaciones estadísticamente significativas no son únicas –en caso contrario, la elaboración del código de reconocimiento sería mucho más sencilla. Como se ve en los ejemplos representados en la Figura 4.6, es frecuente encontrar nucleótidos distintos (normalmente dos) en clara competencia por alguna de las posiciones asociadas a especificidad: NT-4 en el caso de $K_{15}S_{16}$ y NT-5 en el resto de logos de esta Figura[5]. Estas ambigüedades, que casi siempre sobreviven a la purga de las combinaciones no significativas[6], suelen tener en sí mismas un carácter palindrómico, –por ejemplo, si en NT-5_L compiten los nucleótidos A y G, en NT-5_R lo hacen T y C, lo que apunta a que la competencia se establece entre dos combinaciones palindrómicas alternativas para [NT-5_L, NT-5_R], a saber, [A, T] y [G, C].

Para ahondar en esto, fijémonos en el tercero de estos logos, correspondiente a dominios con $I_{15}A_{16}$ y con ambigüedad en NT-5. A simple vista se ve que, al menos formalmente, esta ambigüedad puede ser entendida como la competencia por las posiciones (NT-5_L, NT-4_L; NT-4_R, NT-5_R) de las combinaciones palindrómicas P1=(TA; TA) y P2=(AA; TT) en las que hemos coloreado las posiciones que difieren entre una y otra. A priori, la presencia de esta ambigüedad podría obedecer a distintos escenarios. En uno de ellos, la ambigüedad surge de la acumulación en un mismo logo de BSs asociados a TFs con especificidades excluyentes hacia

[5]El fenómeno de competencia puede darse simultáneamente en NT-4 y NT-5. Por ejemplo, este es el caso de $R_{15}S_{16}$ (véanse los logos del Apéndice D).

[6]No obstante, las posiciones que presentan ambigüedades aportan menos al contenido de información del logo que aquéllas otras en las que claramente domina un nucleótido concreto. Ello provoca que en estos casos sea más fácil que puedan ser aceptadas como BSs secuencias que incorporan en estas posiciones nucleótidos por los que el TF no tiene una afinidad especial, puesto que la penalización que conlleva esta incorporación en la búsqueda mediante PWM es menor que la que resultaría de la sustitución de un nucleótido dominante.

secuencias palindrómicas distintas –esto es, los TFs que reconocieran a P1 no serían capaces de unirse al palíndromo alternativo P2 y viceversa. En este caso, y si consideramos un TF de manera aislada, la ambigüedad del logo es de carácter *extrínseco* puesto que no implica una ambigüedad en la especificidad del TF. En un escenario alternativo, y siguiendo con el ejemplo, todos o la mayoría de los TFs serían capaces de unirse a ambos palíndromos e incluso a las combinaciones mixtas (por tanto, no palindrómicas) de sus semisecuencias: M1=(TA; TT) y M2=(AA; TA). Diríamos entonces que la especificidad de cada TF tiene un cierto grado de ambigüedad *intrínseca*. El primer escenario obligaría a definir códigos de reconocimiento parciales, cada uno de los cuales sólo regiría sobre un subconjunto de los TFs con $I_{15}A_{16}$; en cambio, el segundo permite un código de mayor universalidad, código que tendría que recoger esta ambigüedad intrínseca.

Nótese que las secuencias de M1 y M2 son complementarias –va a suceder lo mismo en cualquier mixtura de dos semisecuencias palindrómicas. Si consideramos la doble hebra de ADN, esto implica simplemente un cambio de orientación del BS respecto al gen regulado. Por tanto, a partir de ahora vamos a asumir que las combinaciones M1 y M2 son equivalentes –en cuanto a la energía de enlace TF/BS se refiere– y las vamos a agrupar en un mismo *estado* M. En la notación de estados, que va entre corchetes, los palíndromos son fácilmente identificables: P1={TA; TA} y P2={AA; AA}. En la misma notación, el estado mixto es M={AA; TA} (véanse los detalles sobre la notación en la sección de *Materiales y métodos*). Debido a la agrupación en un mismo estado de las secuencias que son complementarias, el número total de estados posibles es de 136. La Tabla 4.7 muestra la frecuencia de cada uno de ellos para el caso de los BSs asociados a $I_{15}A_{16}$. Como se puede ver, la frecuencia de M (suma de las de M1 y M2) supera a la de palíndromo P2, lo que indica que en este caso estaríamos ante un escenario de ambigüedad intrínseca. Más aún, se ha demostrado que un mutante de LacI con secuencia $I_{15}A_{16}$ es capaz de reprimir eficientemente BSs que contienen uno u otro de los palíndromos considerados [36].

Cabe esperar que la ambigüedad sea de tipo intrínseco o extrínseco dependiendo del alineamiento con el que tratemos. No obstante, lo anterior ya sugiere la manera de anotar sistemáticamente el código. Antes que nada, téngase en cuenta que todo estado no palindrómico puede considerarse formalmente como la combinación de dos estados palindrómicos y, a la inversa, a partir de dos palíndromos distintos siempre podemos considerar la existencia de un estado híbrido. Por tanto, excepto en los casos en los que tengamos una única combinación estadísticamente significativa (y además palindrómica), siempre podemos comparar las frecuencias de cada uno de los componentes de la tríada de estados tipo [P1, M, P2].

Así, definamos las frecuencias respectivas como f_{P1}, f_M y f_{P2}, renombrando los palíndromos si es necesario para que $f_{P1} \geq f_{P2}$. La relación entre estas tres frecuencias podría apuntar a distintos escenarios: en el caso de que $f_{P1} \geq f_M > f_{P2}$ la semisecuencia del segundo palíndromo la encontramos sobre todo en combinacio-

nes asimétricas; podemos considerar que existe entonces un palíndromo dominante que quiebra su simetría mediante la mixtura con un segundo palíndromo preferente. Es posible entonces que la ambigüedad no sea en este caso del todo simétrica en el sentido de que la afinidad por uno de los palíndromos fuese mayor que por la del otro. Entonces, las combinaciones asimétricas implicarían afinidades intermedias. Cabría incluso la posibilidad de que estas afinidades intermedias fueran aún compatibles con la operatividad del BS, mientras que con P2 la regulación fuese ya demasiado laxa. Ello podría explicar los casos en los que uno de los palíndromos es encontrado con bastante menos frecuencia que las combinaciones asimétricas.

En cambio, un dominio de las combinaciones mixtas, $f_M > f_{P1} \geq f_{P2}$, reforzaría un escenario de ambigüedad intrínseca que permite un enlace casi indiferenciado a sitios con combinaciones simétricas o asimétricas de las dos semisecuencias. Incluso, si el dominio de M es muy destacado, podríamos deducir que en ese caso el carácter del enlace TF/BS es esencialmente pseudopalindrómico. Finalmente, si por el contrario las mixturas son infrecuentes frente a uno y otro palíndromo, $f_{P1} \geq f_{P2} > f_M$, ello podría deberse a que tenemos BSs incompatibles, con distintos TFs leyendo palíndromos diferentes (ambigüedad extrínseca).

Los tres estados P1, M y P2 están compuestos a partir de combinaciones puras o mixtas de dos semisecuencias S1 y S2 –por ejemplo, en el caso de $I_{15}A_{16}$ tendríamos S1=TA y S2=AA. Por tanto, en lo que a la construcción del código se refiere, basta con anotar estas dos semisecuencias junto con la frecuencia con la que aparecen cada una de las tres combinaciones.

En general, y en ausencia de información experimental, hemos visto que la abundancia de combinaciones mixtas apunta a la existencia de un código degenerado (ambigüedad intrínseca) y su escasez a un escenario de códigos parciales (ambigüedad extrínseca). Esto viene recogido en el código (Tabla 4.3) mediante símbolos situados entre las semisecuencias según el criterio expuesto en la Tabla 4.2. Los símbolos son autoexplicativos y sugieren las distintas relaciones que, según lo expuesto más arriba, se pueden establecer entre las tres frecuencias.

Nótese que en el caso de $I_{15}A_{16}$ los datos experimentales apuntan a un reconocimiento eficaz de todas las combinaciones puras y mixtas de los palíndromos, y esto a pesar de que la relación entre las frecuencias ($f_{P1} = 32$, $f_M = 21$, $f_{P2} = 4$, Tabla 4.3) señalaría a priori a una relación tipo S1->S2. Por ello, aunque los símbolos de la Tabla 4.2 tratan de sugerir los distintos escenarios que acabamos de exponer, es necesario recalcar que, en última instancia y estrictamente, estos símbolos sólo constatan la relación numérica entre las frecuencias de palíndromos y mixturas. En qué medida estas relaciones se correspondan con dichos escenarios queda como objeto de estudio experimental.

En base a todas las consideraciones anteriores hemos construido el código de naturaleza eminentemente palindrómica de la Tabla 4.3. De manera general, encontramos una acusada tendencia a la degeneración del código sobre el aminoácido AA-15, de manera que es frecuente encontrar semisecuencias repetidas dentro de

relación entre frecuencias	código asociado a $A_{15}A_{16}$
$f_{P1} \geq f_M > f_{P2}$	S1->S2
$f_{P1} > f_M = f_{P2}$	S1->S2
$f_{P1} \geq f_{P2} > f_M$	S1><S2
$f_M > f_{P1} \geq f_{P2}$	S1<>S2

Tabla 4.2: *Criterio de notación del código en función de la relación entre las frecuencias de P1, M y P2. Denominamos siempre P1 a la más frecuente de las dos combinaciones palindrómicas.*

una misma columna. Parece pues que, de cara al reconocimiento de la secuencia de nucleótidos, la variación de este aminoácido, aunque no irrelevante, es menos crítica que la de AA-16. Además, la existencia de tal degeneración –ya apuntada por los modelos mutacionales [36]– entra en concordancia con el muy diferente grado de consenso con el que los estudios estructurales han documentado la interacción de uno y otro aminoácido con estas dos posiciones (Tabla 4.1). La degeneración sobre AA-15 queda expuesta de un modo más patente si se representa el código de manera inversa, es decir, agrupando las secuencias de aminoácidos que leen una misma secuencia de nucleótidos (Tabla 4.4).

AA-15 \ AA-16	A	C	G	I	K	M	N	P	Q	S	T	V
A	AA<>TA:0/6/0			TT><AT:7/0/2 AT<>GT:2/3/0		TT->AT:5/3/1				CA->AA:4/4/0		
E						TT:6				GC:8		
H	AA:2								TG<>TT:3/5/1	CA:2		
I	TA->AA:32/21/4	TA:3			AC->CC:10/3/0		GT<>TA:1/3/0 TG:3			AA><CA:16/7/8	AA:4	
K	AA->GA:19/4/0 AA><TA:19/2/4 TA->GA:4/4/0				AC<>TA:6/12/0	TT->AT:43/8/0				GG->CA:44/4/0 GG->AG:44/4/0 GG><GA:44/2/4 GA<>GT:4/8/0	GG><CA:5/0/2	
L	AA><TA:7/3/6					TT:7			TG:3			
M	AA->TA:10/9/0 CA<>TA:0/4/0											
N					AC:4							
P	AA->CA:13/5/0			AT:5		AT->GT:6/3/0				CA->AA:22/17/3	AA<>TA:3/7/0	
Q											AA:8	
R			GG:3							CA><GG:5/0/3		GA:3
S	TA:5					AT:9			TT:3			
T	TA->CA:29/18/2 TA->AA:29/13/0 TA->AT:29/5/0		GA:3			AT->GT:12/4/0				CA:9		
V	TA->AA:67/26/5 TA->GA:67/14/1		GA:19					CT<>TT:1/2/0		GT:9	AA->GA:9/4/0	
Y									TG<>TT:1/13/0 TG<>CG:1/4/0			

Tabla 4.3: *Código de reconocimiento de* [AA-15, AA-16] *sobre* [NT-5, NT-4]. *La notación sigue los criterios de la Tabla 4.2. Las tríadas de números corresponden a* $f_{P1}/f_M/f_{P2}$ *(Véase el texto). Los casos en los que existe una única combinación estadísticamente significativa y palindrómica aparecen en forma de semisecuencia aislada.*

(a)

NT-5	NT-4 A	NT-4 C	NT-4 G	NT-4 T
A	HA	IK		AI
	KA	NK		PI
	LA			PM
	MA			SM
	PA			TM
	IS			
	IT			
	QT			
	VT			
C	AS			
	HS			
	IS			
	PS			
	RS			
	TS			
	KT			
G	TG	ES	RG	VS
	VG		KS	
	KS		RS	
	RV		KT	
T	IA		IN	AI
	KA		LQ	AM
	LA			EM
	SA			KM
	TA			LM
	VA			SQ
	IC			

(b)

NT-5	NT-4 A	NT-4 C	NT-4 G	NT-4 T
A	AA	KK	KS	TA
	IA			AI
	TA			AM
	VA			KM
	AS			
	PS			
	PT			
C	MA	IK	YQ	VP
	PA			
	TA			
	KS			
G	KA			AI
	VA			PM
	KS			TM
	VT			IN
				KS
T	AA		HQ	VP
	MA		YQ	HQ
	KK			YQ
	IN			
	PT			

Tabla 4.4: *Código de reconocimiento de* [AA-15, AA-16] *sobre* [NT-5, NT-4] *(representación inversa). Se agrupan las secuencias de aminoácidos que reconocen una misma semisecuencia de nucleótidos, distinguiendo los casos en que: a) las semisecuencias aparecen preferentemente en combinaciones palindrómicas (i.e. las que en la Tabla 4.3 se localizan a la izquierda del signo* `->`*, a ambos lados de* `><` *o de manera aislada); b) las semisecuencias con tendencia a la combinación mixta con otras semisecuencias (a la derecha del signo* `->` *o a ambos lados de* `<>`*).*

4.2.5. Equivalentes naturales de los BSs palindrómicos sintéticos de LacI

En la Tabla 4.3 las secuencias de aminoácidos $Y_{15}Q_{16}$ y $H_{15}Q_{16}$ aparecen asociadas al reconocimiento de mixturas de las semisecuencias TG y TT –la ambigüedad (coloreada) reside en este caso en la posición NT-4. Naturalmente, con estas semisecuencias también se pueden construir combinaciones palindrómicas como los dos operadores sintéticos y perfectamente palindrómicos que aparecen en la Tabla 4.5 y que se diferencian entre sí únicamente en las posiciones NT-4_L y NT-4_R (mantenemos en esta Tabla el mismo código de colores asociado a una y otra semisecuencia). El primero de estos palíndromos es SymL, que proviene de la simetrización de la semisecuencia izquierda del BS natural O1 de LacI de *E. coli* [32] y que ya apareció en la Tabla 4.1. Tanto LacI, que tiene precisamente la combinación de aminoácidos de reconocimiento $Y_{15}Q_{16}$, como su mutante $H_{15}Q_{16}$ son capaces de reprimir la transcripción desde este BS [36]. Pero sólo el mutante es capaz de reconocer el segundo de los palíndromos –con etiqueta 344 en el artículo original [36].

Pues bien, existen dominios en el seno de la familia HTH-LacI que son los equivalentes naturales del mutante $H_{15}Q_{16}$. Todos los dominios con esta secuencia pertenecen a distintas especies de actinobacterias y, a pesar de su distancia filogenética con LacI, exhiben un comportamiento coherente con el modelo mutacional. Esto se puede comprobar en la Tabla 4.5, que muestra el conjunto de BSs que hemos encontrado asociados a estos dominios. Efectivamente, todos los BSs están constituidos esencialmente por distintas combinaciones de los dos palíndromos perfectos. En ocasiones, estos BS naturales se acercan mucho a la palindromía perfecta, pero abundan sobre todo los BSs asimétricos en las posiciones NT-4_L y NT-4_R que corresponden a mixturas de ambos palíndromos.

Curiosamente, el BSs natural O1 presenta el mismo tipo de asimetría, aunque desplazada en una base por el intercalamiento mencionado más arriba. Nótese que la delección de NT-2_R en O1 da lugar a una mixtura semejante a las encontrada para los $H_{15}Q_{16}$ (Tabla 4.5). Sin embargo, LacI es incapaz de reprimir la expresión desde un BS como éste [104], lo que apunta a que la asimetría y la inserción de NT-2bis en O1 son fenómenos acoplados. Así, para que exista regulación, la eliminación del inserto ha de ir acompañada de la sustitución en NT-4_R de la adenina por una citosina, lo que da lugar a un operador tipo SymL.

Pero, ¿existe una versión natural de la regulación de SymL por LacI o es ésta un mero artefacto de laboratorio [37]? La mayoría de los dominios naturales con $Y_{15}Q_{16}$ se unen a sitios similares a O1. No obstante, en la Tabla 4.5 documentamos tres casos de BSs que, asociados a secuencias con $Y_{15}Q_{16}$, carecen del nucleótido adicional y de la asimetría en NT-4. En cambio, los tres tienen secuencias muy cercanas a las del palíndromo SymL, con lo que el fenómeno de regulación sobre esta construcción sintética tendría su correspondencia natural.

BS natural O1 para LacI de *E. coli*

AAT TGTGAGC·GGATAACA ATT

BS sintético: O1 con delección de NT-2_R [104]

AAT TGTGAGC·GATAACA ATT

BSs palindrómicos sintéticos de [36]

BS ID [36]	secuencia BS	mutante LacI de *E. coli*
310 (SymL)	AAT TGTGAGC·GCTCACA ATT	$Y_{15}Q_{16}$ (wt) y $H_{15}Q_{16}$
344	AAT TGTTAGC·GCTAACA ATT	$H_{15}Q_{16}$

BSs naturales para $H_{15}Q_{16}$ (este trabajo)

TF VIMSS-ID	secuencia BS	posición	AU/UD	especie
1810447	AgT TGTGAGC·GCTAACA ATT	-86	UD	AAur
3749792	tga TGTGAcC·GtTCACA tcg	-64	UD	CMM
2169839	gga TGTGAGC·GCTCACA cTc	-77	UD	Noca
2169839	tca TGTGAaC·GCTCACA ATT	-105	UD	Noca
2004131	ggT TGTTAGC·GCTAACt gaa	-65	AU	BAD
2004131	AAT aGTGAGC·GtTAACA gTg	-59	UD	BAD
2004131	AAc aGTGAGC·GCTAACA tTc	-64	UD	BAD
2268667	tAg TGTGAGC·GCTAACg caa	-41	AU	Blon
2268667	AAT TGTGAGC·GCTAACA ccc	-70	UD	Blon

BSs naturales tipo SymL para $Y_{15}Q_{16}$ (este trabajo)

3571242	tAg TGTGAcC·GaTCACA gca	-168	AU	Rcas (cloroflexi)
3654531	tgc TGTGAaC·GtTCACg tga	-39	UD	SACE (actino.)
3760850	gAT TGTGAGC·GtTCACA tgg	-124	AU	sce (δ-proteo.)

Tabla 4.5: *Operadores naturales y sintéticos para $Y_{15}Q_{16}$ y $H_{15}Q_{16}$. Las semisecuencias están separadas por un punto. También hemos separado los tres nucleótidos flanqueantes a cada lado de los BSs, que presentan por lo general un menor grado de conservación. Posiciones relativas al codón de inicio del primer gen regulado corriente abajo. Abreviaturas:* AAur, *Arthrobacter aurescens TC1;* BAD, *Bifidobacterium adolescentis ATCC 15703;* Noca, *Nocardioides sp. JS614;* Blon, *Bifidobacterium longum DJO10A;* CMM, *Clavibacter michiganensis subsp. michiganensis NCPPB 382;* Rcas, *Roseiflexus castenholzii DSM 13941;* SACE, *Saccharopolyspora erythraea NRRL 2338;* sce, *Sorangium cellulosum 'So ce 56'.*

4.3. Materiales y métodos

Selección inicial de las secuencias de los dominios. Las 5597 secuencias correspondientes a dominios HTH-LacI (etiqueta de dominio Smart: SM00354) proceden de la base de datos MicrobesOnline [98]. El dominio SM00354 incluye tanto la región HTH como la hélice-bisagra (*hinge-helix*) que contacta con la zona central del BS (Figura 4.1). La longitud de este dominio tiene un valor medio de 71 aminoácidos y una desviación estándar de 3.5 aminoácidos. Bajo la premisa de garantizar en la medida de lo posible la funcionalidad de los dominios, seleccionamos aquellas secuencias comprendidas en el rango de 71 ± 7 aminoácidos –lo cual abarca al 95 % de las secuencias iniciales, y descartamos las que carecieran de un dominio completo PF00356 –esta etiqueta Pfam corresponde a un dominio de 26 residuos que, a diferencia del anterior, sólo incluye la región HTH. También descartamos tres casos de proteínas con dos dominios SM00354. Al final de este proceso contábamos con 4764 secuencias (el 85 % de las de partida).

Selección de dominios no redundantes. El método de las huellas digitales filogenéticas (*phylogenetic footprinting*, PF) exige que exista una mínima distancia evolutiva entre las especies que se comparan, de tal modo que la mayor conservación de los BSs frente al resto de la zona no codificante se ponga de manifiesto. Por ello, eliminamos las redundancias usando Blasclust, que agrupó las 4764 dominios en conjuntos (*clusters*) de alta similitud de secuencia (umbral de similitud $S = 98$, mínima longitud de cobertura $L = 0.9$ y resto de parámetros fijados por defecto). Tras este proceso seleccionamos una sola secuencia de cada uno de los agrupamientos (2639 secuencias).

Selección de las regiones de búsqueda de los BS. MicrobesOnline incluye predicciones de los genes que forman parte de un mismo operón. Esto permitió que para cada uno de los 2639 dominios seleccionados pudiésemos obtener de esta base de datos las secuencias de nucleótidos siguientes (véase la Figura 4.2): i) la región no codificante hasta un límite de 200 nucleótidos localizada corriente arriba del operón al que pertenece el gen con el dominio HTH-LacI; la secuencia de la región era truncada antes de este límite si se alcanzaba la zona de codificación del siguiente operón situado corriente arriba; ii) una región análoga situada delante del operón localizado corriente abajo siempre y cuando éste estuviese codificado en la misma hebra del operón regulador, lo cual sucede para 1462 de los 2639 dominios (55 %). Al final de este proceso teníamos pues una o dos regiones no codificantes asociadas a cada uno de los dominios. Además, de todas ellas obtuvimos versiones extendidas de 250 nucleótidos que incluyen los que ocupan las posiciones +1 a +50 de la zona codificante corriente abajo. Estas secuencias extendidas no fueron truncadas en ningún caso.

Alineamiento de las secuencias de los dominios. Los dominios HTH-LacI fueron alineados usando Muscle [105]; en concreto, aprovechamos la opción que este programa tiene de añadir secuencias a un alineamiento ya existente, lo cual evita en gran medida los problemas asociados a la pérdida de eficiencia que sufren los algoritmos de alineamiento cuando han de tratar con miles de secuencias. Así, nuestro punto de partida fue el alineamiento de familia Smart que, optimizado manualmente, involucra a 49 dominios SM00354 [106]. De este alineamiento eliminamos aquellas posiciones que mostraban huecos (*gaps*) en más del 80 % de las secuencias, lo que resultó en un *alineamiento-semilla* de 71 posiciones. A continuación, para cada una de las 2639 secuencias aplicamos el siguiente protocolo: i) se añade la secuencia al alineamiento-semilla usando la opción mencionada de Muscle, ii) se descartan todas aquellas posiciones de la secuencia que implican la inserción de un hueco en el alineamiento-semilla , y iii) se retira del alineamiento-semilla la secuencia así depurada y se guarda. Al final del proceso se obtiene un alineamiento para las 2639 secuencias que consta de 71 posiciones (Figura 4.3), ninguna de las cuales exhibe más de un 5 % de huecos.

Agrupación de las secuencias asociadas a una misma secuencia de aminoácidos de reconocimiento y primera búsqueda de motivos por PF. Tras el alineamiento se estudió el patrón de conservación de los residuos y se extrajeron las secuencias de las HRs correspondientes[7]. 1490 de los 2639 dominios presentaban la secuencia TVSR en el cuarteto de residuos [AA-17, AA-18, AA-19, AA-20] (Figura 4.4). Dentro de este subconjunto de dominios se agruparon todas las secuencias intergénicas (en la versión truncada) asociadas a aquéllos con una misma secuencia en [AA-15, AA-16]. Sobre cada uno de estos grupos se procedió a una primera búsqueda de BSs mediante la técnica del PF implementada en el programa Gibbs Motif Sampler [107]. Los parámetros del programa se fijaron como sigue: máximo número de BSs permitidos en cada secuencia: 1; número total estimado de BSs para cada grupo de secuencias no codificantes igual al número de éstas; BSs de tipo palindrómico con una extensión de 14 NT y sin fragmentación. Los resultados eran robustos ante cambios en los parámetros anteriores, incluido el tamaño de los BSs. Cada búsqueda incorporaba la información sobre el modelo de la secuencia de fondo en ausencia de BSs (*background*), también calculado con este mismo programa [107]. En cambio, para evitar problemas de circularidad se evitó el uso de información previa basada en BSs ya documentados. De los sitios encontrados sólo retuvimos aquéllos a los que se les adjudicaba una probabilidad superior al 40 % de ser verdaderos

[7] En toda esta sección y para abreviar designaremos como HR al conjunto de los seis aminoácidos de la hélice de reconocimiento que ocupa el tramo de posiciones críticas que va desde AA-15 a AA-20.

BSs. El uso en este caso de las regiones truncadas (es decir, estrictamente intergénicas) tenía por objeto evitar ejecutar el PF sobre regiones codificantes que, como sucede con los BSs, tienen un grado de conservación superior al que en promedio se da para las regiones intergénicas.

Segunda búsqueda de motivos con la matriz de peso de las posiciones. Todo lo que sigue se aplicó sucesivamente a cada uno de los conjuntos de secuencias asociados a una misma HR: en primer lugar, a partir de los BSs obtenidos en la primera búsqueda se construyó la matriz de peso de las posiciones (*position weight matrix*, PWM) tal y como se cuenta en [29] usando un pseudo-recuento (*pseudocount*) igual a 0.5. A continuación, procedimos a realizar una segunda búsqueda de BSs deslizando esta PWM sobre la versión extendida de las regiones intergénicas, reteniendo en un primer momento todos los sitios con una puntuación (*score*) igual o superior a la mínima encontrada en el conjunto de BSs de partida. Esta segunda selección constituía el conjunto de sitios candidatos de los que retendríamos sólo aquéllos que superasen un cierto umbral de significatividad estadística. Con este fin, a todos estos sitios se les calculó el p-valor en relación a 10^7 puntuaciones obtenidas a partir de versiones aleatorias de las regiones intergénicas, tras lo cual aplicamos el siguiente protocolo de manera iterativa: i) selección de los BSs de p-valor inferior a 10^{-5}; ii) construcción a partir de esta selección de una nueva PWM; ii) puntuación de todos sitios candidatos bajo la nueva PWM; iii) selección de los BSs con un p-valor inferior a 10^{-5} bajo la nueva PWM. El proceso se itera hasta que el conjunto de sitios seleccionados da lugar a una PWM que selecciona a este mismo conjunto de sitios. Con esta selección final, que supone la culminación del proceso de búsqueda de BSs sobre las secuencias correspondientes y que implica siempre a sitios con Z-scores por encima de $Z = 4$, construimos los logos asociados a cada HR. Todos los BSs están leídos en la hebra *sentido* del operón regulado, lo cual implica un etiquetado unívoco para la semisecuencia izquierda y derecha de cada BSs.

Comparación con la RegTransBase. La RegTransBase (RTB) es actualmente la base de datos más completa de BSs procariotas [101]. Los BSs están agrupados en alineamientos que se asocian a TFs ortólogos y a genes regulados también ortólogos. La confección de estos alineamientos parte en ocasiones de regulaciones comprobadas experimentalmente en alguno de los organismos considerados; en otros casos, los alineamientos se infieren computacionalmente *de novo*. De cada uno de los alineamientos se extrae la correspondiente PWM, a la que acompañan indicaciones sobre los taxones bacterianos sobre los se aconseja su aplicación y una representación gráfica en forma de logo.

La RTB no asigna directamente un TF a cada BS, sino que para cada alineamiento proporciona una relación de los TFs ortólogos por un lado y de los BSs por otro. Y no siempre existe un sólo TF ortólogo por organismo. Además, todos los TFs de la relación de ortólogos considerados en un alineamiento tampoco tienen por qué compartir una misma secuencia de los aminoácidos de reconocimiento, puesto que el criterio de ortología se basa en la semejanza del conjunto de la secuencia del TF.

Por tanto, la comparación de nuestros resultados –que clasifica los BSs en función de la HR que los reconoce– con los alineamientos de la RTB exigía un tratamiento previo de éstos. Así que sobre cada uno de los alineamientos que involucran a miembros de la familia de HTH-LacI se procedió como sigue: i) de la relación de TF ortólogos se asignaron como potenciales reguladores de un BS dado aquéllos codificados en el mismo genoma del BS. Con frecuencia (pero no siempre) es uno sólo el TF asignado; ii) se depuraron de los alineamientos todos los BSs asociados a TFs con dominios HTH-LacI redundantes aplicando el mismo criterio de redundancia que empleamos en el tratamiento de nuestros datos (ver más arriba); iii) también se depuró la redundancia asociada a la doble anotación de BSs que regulan a la vez a dos operones orientados divergentemente; y iv) se dividió cada alineamiento en tantos otros como HR distintas estén involucradas en el mismo. De esta manera obtuvimos un conjunto de alineamientos de BSs que, como sucede con los nuestros, estarían vinculados a una misma HR.

Selección de las secuencias significativas para NT-4 y NT-5. Antes que nada, vamos a establecer unos criterios de notación para estas posiciones. Para ello, fijémonos en la Tabla 4.6; en concreto, en la secuencia de doble hebra #1. Supongamos, además, que la hebra superior es la misma correspondiente a la secuencia sentido del gen regulado. Por tanto, esta podría ser una de las típicas secuencias de hebra simple resultantes de nuestra búsqueda de BSs. La lectura en esta secuencia del cuarteto de nucleótidos que ocupan las posiciones NT-4 y NT-5 de ambas semisecuencias la vamos a anotar entre paréntesis, es decir,

$$(\text{NT-}5_L,\ \text{NT-}4_L;\ \text{NT-}4_R,\ \text{NT-}5_R)\rightarrow(\text{GT};\ \text{CT}).$$

Como se verá en seguida, resulta conveniente anotar las posiciones correspondientes a la semisecuencia derecha leídas en la hebra antisentido (y además desde la posición 5' a la 3' de esta hebra). Esta vez, la notación hace uso de los corchetes (c indica el nucleótido complementario) :

$$[\text{NT-}5_L,\ \text{NT-}4_L;\ \text{NT-}5^c_R,\ \text{NT-}4^c_R]\rightarrow[\text{GT};\ \text{AG}].$$

Obsérvese que los palíndromos –como las secuencias #3 y #4 en la Tabla 4.6– se identifican sin esfuerzo con esta última notación:

secuencia #1
5’-tgGTa-c·g-tCTca-3’
3’-acCAt-g·c-aGAgt-5’
secuencia #2
5’-tgAGa-c·g-tACca-3’
3’-acTCt-g·c-aTGgt-5’
secuencia #3
5’-tgAGa-c·g-tCTca-3’
3’-acTCt-g·c-aGAgt-5’
secuencia #4
5’-tgGTa-c·g-tACca-3’
3’-acCAt-g·c-aTGgt-5’

Tabla 4.6: *Ejemplos de BSs que incluyen las dos hebras. Las secuencias no palindrómicas #1 y #2 son exactamente iguales salvo cambio de orientación. La secuencias #3 y #4 corresponden a los palíndromos construidos a partir de las semisecuencias de #1 y #2. Coloreamos posiciones NT-4 y NT-5 para ayudar a la visualización. Obsérvese que en el resto de posiciones hemos colocado arbitrariamente los nucleótidos dominantes en el logo consenso (Figura 4.7).*

$$(\mathrm{AG};\,\mathrm{CT})=[\mathrm{AG};\,\mathrm{AG}] \qquad (\mathrm{GT};\,\mathrm{AC})=[\mathrm{GT};\,\mathrm{GT}].$$

Existen $16^2 = 256$ combinaciones de este tipo. La frecuencia $\tilde{f}$ de aparición de cada una de ellas en un alineamiento de BSs dado puede disponerse en una matriz de elementos $\tilde{f}_{i,j}$ donde los dos subíndices recorren las 16 combinaciones posibles de dos nucleótidos. Las frecuencias correspondientes a secuencias palindrómicas ocuparían la diagonal principal, $\tilde{f}_{i,i}$; y las combinaciones que son invariantes salvo en su orientación, como #1 y #2, intercambiarían los índices de filas y columnas. La suma de todas las frecuencias es el número N de BSs del alineamiento.

Precisamente, nosotros hemos considerado equivalentes –en cuanto a la energía de enlace TF/BS se refiere– aquellas secuencias que presentan este tipo de invariancia[8]. Por ello, vamos a agrupar estas secuencias en lo que definimos como un mismo *estado*, que anotaremos haciendo uso de llaves:

$$\{\mathrm{AG};\,\mathrm{GT}\}=[\mathrm{AG};\,\mathrm{GT}]\cup[\mathrm{GT};\,\mathrm{AG}]=(\mathrm{AG};\,\mathrm{AC})\cup(\mathrm{GT};\,\mathrm{CT}).$$

Nótese que en el caso palindrómico sólo existe una secuencia por estado:

$$\{\mathrm{AG};\,\mathrm{AG}\}=[\mathrm{AG};\,\mathrm{AG}]=(\mathrm{AG};\,\mathrm{CT}).$$

[8] A pesar de que en este ejemplo con el par #1 y #2 la invariancia implica a todas las posiciones, basta con que se dé para las posiciones NT-4 y NT-5.

Esta va a ser la notación final de nuestro espacio de secuencias, que ahora queda restringido a $16 \times (16+1)/2 = 136$ estados distintos. Por tanto, el conjunto de frecuencias f de cada uno de estos estados se puede agrupar en la submatriz triangular superior (incluida la diagonal para los palíndromos) de una matriz 16 x 16, de manera que $f_{i,j} = \tilde{f}_{i,j} + \tilde{f}_{j,i}$ para los casos no palindrómicos, y $f_{i,i} = \tilde{f}_{i,i}$ para los palindrómicos. Por cada uno de los alineamientos de BSs elaboramos una matriz de frecuencias de este tipo, matriz que denominaremos F. La suma de todas las frecuencias de la matriz vuelve a coincidir con el número N de BSs del alineamiento. La Tabla 4.7 muestra el caso correspondiente a la secuencia de aminoácidos de reconocimiento $\mathrm{I}_{15}\mathrm{A}_{16}$.

	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
AA	4	0	0	0	0	0	0	0	2	0	0	0	21	0	0	0
AC		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
AG			0	0	0	0	0	0	0	0	0	0	0	0	0	0
AT				0	0	0	0	0	0	0	0	0	1	0	0	0
CA					0	0	0	0	0	0	0	0	1	0	0	0
CC						0	0	0	0	0	0	0	0	0	0	0
CG							0	0	0	0	0	0	0	0	0	0
CT								0	0	0	0	0	0	0	0	0
GA									1	0	0	0	1	0	0	0
GC										0	0	0	0	0	0	0
GG											0	0	0	0	0	0
GT												0	0	0	0	0
TA													32	0	0	0
TC														0	0	0
TG															0	0
TT																0

Tabla 4.7: *Ejemplo de matriz F de las frecuencias de los 136 estados posibles (secuencias distintas salvo inversión) para las posiciones NT-4 y NT-5 de ambas semisecuencias. Este caso corresponde a los 63 BSs del alineamiento asociado a los aminoácidos de reconocimiento $I_{15}A_{16}$. Se han resaltado en rojo las frecuencias estadísticamente significativas (véase el texto)*

Por las razones mencionadas en el texto principal, de cara a la elaboración del código se descartaron aquellos estados carentes de un mínimo de significatividad estadística frente a un modelo nulo de distribución aleatoria de las frecuencias. Este modelo se construyó bajo la hipótesis de que la probabilidad de cada estado vendría dada simplemente por la probabilidad de ocurrencia de las secuencias de nucleótidos correspondientes en el *background* intergénico. Es decir, y siguiendo con los ejemplos de más arriba, tendríamos que la probabilidad en el modelo nulo de los estados {AG; GT}

y {AG; AG} vendría dada, respectivamente, por:

$$\begin{aligned} p_{\{AG;\,GT\}} &= p_{[AG;\,GT]} + p_{[GT;\,AG]} = p_{(AG;\,AC)} + p_{(GT;CT)} = \\ &= p_A\, p_G\, p_A\, p_C + p_G\, p_T\, p_C\, p_T \\ p_{\{AG;\,AG\}} &= p_{[AG;\,AG]} = p_{(AG;\,CT)} = p_A\, p_G\, p_C\, p_T \end{aligned}$$

siendo p_i $(i = A, C, G, T)$ la probabilidad de fondo de los cuatro nucleótidos, que está calculada sobre la hebra sentido.

Así, de cara al cálculo del p-valor de cada estado observado, construimos un conjunto de 10^4 matrices F aleatorias, $F^{(rnd)}$, en las que la ocupación de los 136 estados viene determinada por la probabilidad de fondo de los mismos, $p_{\{s\}}$ $(s = 1 \ldots 136)$. Si las probabilidades de los nucleótidos están normalizadas $(\sum_{i=A,C,G,T} p_i = 1)$ también lo estarán las $p_{\{s\}}$.

Algoritmo 4.1 Pseudocódigo para la generación del conjunto de matrices $F^{(rnd)}$

```
for j = 1 to 9999 do
  for i = 1 to N do
    generar u según U(0|1)
    selección de s tal que c_s = máx({c_r|c_r ≤ u})
    f_{s,j}^{(rnd)} ← f_{s,j}^{(rnd)} + 1
  end for
end for
```

La manera de construir el conjunto de matrices $F^{(rnd)}$ se muestra en el Algoritmo 4.1, en el que las c_s $(s = 1 \ldots 136)$ son los puntos de inicio de las particiones en que se divide el segmento unitario [0,1]. Esta partición es proporcional a las probabilidades $p_{\{s\}}$:

$$c_s = \begin{cases} 0 & s = 1 \\ \sum_{r=1}^{s-1} p_{\{r\}} & s = 2 \ldots 136 \end{cases}$$

En el bucle interno del algoritmo se generan N números aleatorios u según la distribución uniforme estándar $U(0, 1)$. Cada vez que se genera un número aleatorio su valor se ubicará en una de las particiones del segmento de probabilidades. El estado asociado a esa partición vería incrementada en una unidad el valor de su frecuencia. El proceso se repite 10^4 veces[9]. Finalmente, para cada uno de los estados s se calcula el p-valor P_s de cada una de las

[9]En el cálculo de p-valores se considera que el conjunto observado cuenta como una realización aleatoria más [108], de ahí que, en rigor, se generen exactamente 9999 matrices $F^{(rnd)}$. El 1 que aparece en el numerador de la expresión del p-valor obedece a esto mismo.

frecuencias observadas f_s frente al conjunto de frecuencias aleatorias para ese mismo estado $f_{s,j}^{(rnd)}$ $(j = 1 \ldots 9999)$,

$$P_s = \frac{(\text{número de } f_{s,j}^{(rnd)} \geq f_s) + 1}{10^4}$$

En la matriz de la Tabla 4.7 se han resaltado en rojo las frecuencias correspondientes a $P_s < 0.05$ –una vez corregidos del proceso de comparación múltiple (*multiple testing*) [85]. Los estados asociados a estos p-valores significativos son la base de la anotación del código de reconocimiento de la Tabla 4.3. Los estados no diagonales siempre pueden considerarse formalmente como derivados de la combinación de las semisecuencias de dos estados palindrómicos, lo que da pie a que puedan compararse las frecuencias de la tríada (P1, M, P2) según el criterio de la Tabla 4.2.

En todo caso, para poder establecer estas comparaciones se consideraron las siguientes frecuencias, aun cuando no fuesen estadísticamente significativas: i) las frecuencias de los dos palíndromos asociados a una combinación no palindrómica significativa, y ii) la frecuencia del estado no palindrómico correspondiente a la mixtura de dos palíndromos significativos. Por último, en la Tabla 4.3 aparecen elementos aislados (no forman parte de tríadas) que corresponden a alineamientos que sólo cuentan con una única secuencia significativa que además es palindrómica.

Capítulo 5

Conclusiones

Esta tesis recoge los resultados de tres trabajos que, en torno a la autorregulación y a los motivos de red, abarcan muy distintos niveles de organización biológica. Inspirados quizás por la estrategia que va de lo pequeño a lo grande en una de las aproximaciones posibles a la Biología de Sistemas, la aproximación *bottom-up*, el foco de nuestras investigaciones se ha ido ampliando desde el estudio de un sistema concreto, que a lo sumo implica a unas pocas especies moleculares distintas, para, tras una parada intermedia en la red transcripcional completa de *E. coli*, llegar a lidiar con todos los genomas procariotas secuenciados a día de hoy. Este orden creciente en la magnitud de los sistemas considerados no es casual, puesto que, tras cada trabajo, siempre cabía la duda de hasta qué punto los resultados obtenidos eran generalizables: del sistema/motivo SOS al resto de motivos de *E. coli* y luego de este organismo modelo a todos los procariotas.

En primer lugar, nuestro tratamiento matemático-experimental del sistema SOS de *E. coli* como modelo genético válido para estudiar la estructura de control en un sistema natural, reveló la optimización que el control autógeno de su regulador maestro supone para la respuesta del sistema SOS ante daños en el ADN. Efectivamente, en contraste con una construcción sintética alternativa sin autorregular, el sistema natural exhibía i) una mayor robustez del estado estacionario en el que se encuentra el sistema en ausencia de daños –robustez muy cercana a su máximo teórico–, lo que prevenía el disparo innecesario de la transcripción de los genes SOS; ii) menores tiempos de recuperación desde el estado inducido al estado estacionario original; y iii) una respuesta mucho más graduada a las distintas dosis de daño infligidos en el ADN y que se traducía en descensos de la tasa de crecimiento menos drásticos. En comparación, el sistema sin control autógeno presentaba siempre sobreinducción en relación al autorregulado, lo cual implica un alto coste de la respuesta del sistema –con los consiguientes efectos notorios en las curvas de crecimiento– incluso ante pequeñas dosis de daño.

En cuanto a los tiempos de respuesta en el trayecto de ida desde el estado sin inducir al de máxima inducción, los dos circuitos autorregulado y sin autorregular

presentaban un comportamiento muy semejante. Esto no sería precisamente lo esperable si el circuito implicara sólo conexiones transcripcionales –como el tratado en el Capítulo 1 y en el que una autorregulación más fuerte siempre comporta mayor velocidad a la hora de alcanzar un nuevo estado estacionario [8, 53]. La discrepancia con lo observado en el SOS se debe a que la dinámica de éste justo tras el momento en que acaece el daño está presumiblemente dominada por el proceso postraduccional de autoproteólisis de LexA mediada por RecA* [75], lo cual supone de hecho una llamada a la cautela frente a los intentos de deducir directamente una dinámica partiendo tan sólo del esquema de conexiones transcripcionales de los motivos [13]. En cualquier caso, otras propiedades asociadas previamente a la autorregulación (robustez del estacionario, moderación de la respuesta) [8] y comprobadas experimentalmente mediante la construcción de pequeños circuitos sintéticos [53,58], fueron observadas y cuantificadas por primera vez en un sistema natural. Incluso documentamos la mayor aceleración de la respuesta en el camino de vuelta al estacionario bajo condiciones de bajas dosis de daño, condiciones en las que parecen dominar los tiempos transcripcionales de la autorregulación.

Las conexiones transcripcionales del sistema SOS dibujan un arquetípico motivo SIM, en el que un TF regula la transcripción de un conjunto de operones de manera exclusiva, es decir, sin que concurran sobre éstos regulaciones adicionales provenientes de otros TFs. Los resultados previos contribuyen a la caracterización de las ventajas funcionales de la regulación autógena en sistemas inducibles controlados por un represor maestro. Cuando la respuesta en estos sistemas lleva parejo un coste en el crecimiento de las bacterias, como en el sistema SOS, la regulación precisa asociada a la estrategia autógena minimiza tal coste. Probablemente, la presencia mayoritaria de la autorregulación en otras redes que exhiben una arquitectura tipo SIM [9, 10] obedezca a la misma razón de evitar una respuesta excesiva cuyo coste se vería amplificado por toda la batería de genes regulados.

***

La constatación del fenómeno de la optimización de este sistema genético natural por vía del modo de control autógeno de su regulador maestro, y el hecho de que este sistema genético exhibiese el esquema transcripcional de uno de los motivos de red, suscitaba la posibilidad de racionalizar a través de los motivos el papel dinámico de este tipo de control en el conjunto de la red, estudiando la manera en la que la autorregulación se integra en estas arquitecturas intermedias. En este sentido, y aparte del existente con los reguladores maestros de los SIM, ya se había apuntado de manera cualitativa el vínculo de la autorregulación con los elementos Y de los FFLs [10].

Evidentemente, conforme se abordan niveles de organización biológica de mayor envergadura se va perdiendo resolución en el detalle con el que los sistemas pueden ser resueltos, lo que cambia en última instancia el tipo de preguntas que

se pueden plantear sobre los mismos. Por ello, del tratamiento matemático de las interacciones moleculares del SOS y de las medidas de alta resolución de fluorimetría y luminometría, que eran capaces de explicar con bastante precisión el fenotipo del sistema, hemos pasado, cuando se trata de generalizar los resultados al conjunto de motivos de *E. coli*, a un tratamiento de tipo estadístico de la red de transcripción de esta bacteria.

No obstante, el éxito en el tratamiento de un sistema de funcionalidad tan definida como el SOS no debía dar lugar a una generalización a la ligera en el ámbito de los motivos de red, pues existe una cuestión que ha de tenerse en cuenta en todo momento y que no es sino el debate sobre el propio carácter funcional que se les adjudica a éstos. Esta adjudicación se basa tanto en la significatividad de su sobrerrepresentación frente a modelos nulos de la red de transcripción, que se considera indicio de selección adaptativa, como en las propiedades de procesamiento de información que se les atribuye a los motivos o sus agregados [10,17–19]. Como hemos visto, existen contraargumentos que critican tanto el origen adaptativo de los motivos [15, 77–79] como el carácter determinante que, a la hora de inferir una funcionalidad, tiene su estructura de conexiones [13, 14, 16, 70]. A esto hay que añadir la cuestión de la propia significatividad estadística de cada uno de los motivos por separado, que no había sido considerada previamente en los trabajos basados en recuentos globales de los mismos [9, 61, 62].

Cubriendo este último hueco, nuestros resultados plantean la posibilidad de un escenario intermedio que reconcilia parcialmente las posturas neutrales y adaptativas. Así, no se puede descartar –al menos en términos estadísticos– que existan motivos (o agregados de motivos) seleccionados efectivamente como entes funcionales (lo que se manifestaría en una fuerte señal estadística individual) en convivencia con otros que no serían más que un efecto colateral del establecimiento de las regulaciones que los constituyen. En este sentido, cuando se distinguen los motivos uno a uno, la subred del SOS destaca por su fuerte significatividad estadística individual frente a modelos nulos de redes aleatorias, y esto aún dentro de los motivos SIM, cuya definición implica de partida para estas arquitecturas un notable grado de aislamiento respecto al resto de la red transcripcional. Realmente, sólo unos pocos TFs dan lugar a motivos SIM significativos; se trata siempre de TFs de grandes regulones sobre los que el control transcripcional se ejerce con una exclusividad muy acusada (alta *SIMness*). El significativo aislamiento de una subred que involucra a dos decenas de operones hace del sistema SOS un ejemplo particularmente diáfano de la identificación de un motivo de red, en este caso un motivo SIM, con un módulo funcional autónomo, identificación que, ni de lejos, está tan clara en el resto de motivos.

Antes de considerar el papel que en la integración del control autógeno en la red transcripcional tienen los motivos (particularmente el FFL, el más estudiado de los motivos de red, con la excepción quizás de la propia autorregulación), realizamos un primer mapeado de esta integración clasificando los TFs de la red en

seis categorías: en primer lugar, distinguiendo entre aquéllos con o sin regulación transcripcional externa, lo que equivalía a reducir la estructura multicapa de la red de transcripción a dos niveles: el superior, constituido por los TFs carentes de regulación externa; y el inferior, en el que se engloban todos los demás. En segundo lugar, dentro de cada uno de los dos niveles clasificamos los TFs en tres clases de conectividad (baja, media y alta) basadas en el número de operones regulados.

Al estudiar los porcentajes de TFs autorregulados en cada una de las seis categorías, el nivel superior en general mostraba una menor incidencia del control autógeno, excepto en el caso de los TFs de alta conectividad (*hubs*). Es probable que esta excepcionalidad se deba a la necesidad de un control adecuado sobre estos TFs globales, capaces de producir grandes cambios en la fisiología celular. Por otro lado, la mayor presencia de autorregulaciones en promedio en el nivel inferior podría deberse al mencionado vínculo del control autógeno con los elementos Y de los FFLs.

Para cuantificar este vínculo de una manera que eludiera las redundancias derivadas de la formación de agregados de FFLs, se definió sobre cada TF con regulación externa una magnitud (la *FFLness*) que daba cuenta de la tendencia del TF a formar parte de FFLs en el rol de elemento Y. La medida promedio de esta magnitud en las tres clases de conectividad mostraba no sólo que el vínculo de la autorregulación con los Y's está sesgado hacia la clase de baja conectividad sino que la propia constitución de FFLs es especialmente significativa entre los miembros de esta misma clase. La tendencia al ensamblaje de FFLs, por el contrario, mengua con la conectividad del TF hasta situarse en el caso de los *hubs* poco por encima de lo esperable de manera neutral. Esto último se debe a que muchas de las corregulaciones establecidas por parejas de estos TFs globales dan lugar al ensamblaje de un número considerable de FFLs (y de bi-fanes) explicables bajo un modelo neutro de solapamiento aleatorio de grandes regulones. No obstante, algunas de estas corregulaciones (y, con ellas, los motivos de red generados) tenían una fuerte significatividad.

Siguiendo con los FFLs con Y de baja conectividad –que, en función de lo anterior, parecen soportar buena parte de la señal estadística global de los FFLs–, tratamos de afianzar, más allá de los argumentos estadísticos, el escenario adaptativo para este subconjunto de motivos. Para ello, realizamos una doble criba que descartara provisionalmente este escenario para todo aquel FFL en el que concurriese alguna circunstancia que pudiese explicar su ensamblaje por mecanismos neutrales. Las circunstancias consideradas fueron, por un lado, la arquitectura genómica (en concreto, la localización adyacente del operón que codifica al TF que hace de Y y uno de sus pocos operones Z regulados); y por otro, los fenómenos de homologías entre miembros del FFL, de modo que se pudiese atribuir –aunque fuese de manera remota– el origen del FFL a una duplicación ancestral de alguno de sus constituyentes. A pesar de que la criba fue realizada con severidad, una

fracción no desdeñable de este conjunto de FFLs se resistía a una explicación neutral; máxime cuando la mayoría de los supervivientes a la criba compartían atributos funcionales claramente ligados a la lógica dual y jerárquica de la represión catabólica –en la que concurren un TF global (CRP) en el papel de X y un TF específico en el rol de Y. Esto último era extensible a muchos de los FFLs que habían sido etiquetados provisionalmente como neutrales, lo que obligaba a reconsiderar para ellos un origen de tipo adaptativo. Y es que, en conjunto, unos y otros FFLs adoptaban la forma de agregados claramente identificables como unidades funcionales autónomas.

La autorregulación del elemento Y implica que la lógica dual afecta también al conjunto de genes codificados en el operón que también codifica a Y. Además de implementada en forma de FFL, la lógica dual aparece con frecuencia en forma de policistrón autorregulado y con regulación externa adicional, lo que apunta a que es precisamente esta lógica jerárquica el objeto último de selección. La autorregulación del TF específico no es un requisito necesario para el funcionamiento de la represión catabólica, como demuestra el paradigmático caso de LacI, aunque tal autorregulación puede llegar a ser ventajosa. Esta contingencia en la presencia/ausencia del control autógeno puede ser la razón de que el sesgo hacia la autorregulación de los Y de baja conectividad no termine de ser tan significativo como el propio ensamblaje de estos agregados de FFLs *jerárquicos*.

***

El segundo salto de escala en el desarrollo de esta tesis lo supone el que va desde el organismo modelo *Escherichia coli* al resto de organismos procariotas secuenciados. La pérdida de resolución en la descripción funcional que se produce al pasar de un sistema a un organismo se agudiza cuando se trata de lidiar con centenares de genomas, en los que las redes de transcripción están sólo parcialmente determinadas –muchas veces sólo gracias a la búsqueda, no exenta de dificultades, de ortólogos de secuencias de proteínas con relaciones de regulación conocidas. Como compensación, una de las grandes virtudes de la genómica comparada estriba en arrojar luz sobre la cuestión de qué hay de genérico y qué de particular en las propiedades caracterizadas en los organismos modelo.

La regulación adyacente, sobre todo la que se produce con arquitecturas divergentes entre el operón regulado y aquél que codifica el regulador, está vinculada a la autorregulación, puesto que una y otra se realizan desde una zona intergénica común, lo cual propicia además la implementación ya mencionada de las lógicas duales en forma de FFL. Además, si a esto unimos que este tipo de adyacencia puede dar lugar a la transferencia horizontal de grupos de genes que constituyen una funcionalidad autónoma [90] o, en el ámbito intracelular, a la localización eficaz de BSs por parte de su regulador correspondiente [28], todo apunta a que la regulación adyacente es sin duda objeto de selección adaptativa.

Así, este segundo salto del que hablamos viene propiciado por la asunción de que la autorregulación y el fenómeno de regulación adyacente en general constituyen fenómenos universales que abarcan a todos o la mayoría de procariotas. Esta asunción, respaldada además por la predicción de los regulones que se viene realizando sobre cada genoma procariota recién secuenciado [98], da pie a emprender una búsqueda local de BSs en las zonas no codificantes situadas en el entorno del operón que codifica al TF que reconoce tales sitios. La eficiencia de esta estrategia de búsqueda, además de añadir argumentos en pro de la bondad de la asunción en la que descansa, permitió, en último término, la definición de un código de reconocimiento proteína/ADN dentro de la amplia familia de reguladores que exhiben el dominio HTH-LacI, dominio responsable de la unión al BS. La búsqueda de este código descansaba, además, en la hipótesis de que el contexto que para los aminoácidos que contactan directamente el ADN supone la propia estructura de la proteína no variaría tanto dentro de esta familia como para impedir la formulación de unas reglas de reconocimiento que rigieran sobre una fracción importante de dicha familia [39].

Encontrar un código de reconocimiento de la mayor universalidad posible requiere que la búsqueda de BSs se extienda a todo genoma procariota secuenciado, una extensión para la cual el método de búsqueda de BSs basado en la regulación adyacente proporciona las siguientes ventajas: i) no se depende de la identificación y el agrupamiento de ortólogos entre los genes regulados; en cuanto a los reguladores, basta con que todos pertenezcan a la familia HTH-LacI y que compartan una misma hélice de reconocimiento; ii) en el caso de la autorregulación se elude inmediatamente la cuestión de si existe relación funcional entre el TF y el operón regulado; y en el caso de la adyacencia, es de esperar que cuando no exista relación funcional entre los operones vecinos, la vecindad no se conserve más allá de las pequeñas distancias evolutivas; iii) se eliminan muchos de los problemas de los métodos basados en la extrapolación de relaciones funcionales conocidas a otros organismos distintos de aquél en el que fueron originalmente descritas, incluyendo los sesgos hacia las funciones presentes en los organismos modelo de partida; iv) la funcionalidad deja de ser un aspecto apriorístico y puede rescatarse a posteriori como elemento de análisis de las regulaciones encontradas sobre los operones vecinos; v) se elimina asimismo el problema de las duplicaciones o supresiones de ortólogos de los TFs o los operones regulados, puesto que ahora existe una correspondencia unívoca entre el TF y las regiones donde se buscan los BSs. La selección de estas regiones se convierte en un procedimiento automático en todo genoma recién anotado, independientemente de su localización filogenética. Como consecuencia de esto, el número de dominios a los que se les puede asociar al menos un BS casi triplica al que se obtiene a partir de las bases de datos más completas basadas en la ortología. A la hora del recuento final del número de BSs asociados a una misma secuencia de hélice de reconocimiento, esta riqueza en el número de TFs abarcado es capaz de compensar la restricción de la búsqueda de

BSs al entorno de la región codificadora del TF. Por supuesto, nuestro método no sustituye al tradicional, sino que debe verse como un enfoque complementario a éste en la resolución de las redes de transcripción bacterianas y que, además, resulta particularmente apropiado cuando se trata de desvelar códigos como el que nos ocupa.

A partir de las secuencias consenso de los BSs asociados a unos mismos aminoácidos de reconocimiento AA-15 y AA-16 construimos un logo-consenso que muestra una clara delimitación entre posiciones muy poco conservadas y otras con un alto grado de conservación: entre estas últimas se encuentra el par [G, C] en [NT-6_L, NT-6_R], cuya conservación va presumiblemente asociada a su enlace por parte del residuo R_{20} que incluyen todos los dominios HTH-LacI que hemos considerado; la especificidad asociada a este enlace vería así confirmada su independencia de actuación respecto de la vinculada a los contactos que se realizan desde el par de aminoácidos [AA-15, AA-16] sobre NT-4 y NT-5 [37]. El logo-consenso de todos los BSs presenta en estas posiciones una alta variabilidad que se corresponde con la que se observa para AA-15 y AA-16 en el alineamiento de los dominios HTH-LacI. En conjunto, todo esto permitió la definición de un código de reconocimiento coherente basado únicamente en la correspondencia entre estas posiciones de aminoácidos y nucleótidos.

En la anotación del código hemos tenido en cuenta que las ambigüedades observadas en las posiciones NT-4 y NT-5 pueden obedecer a distintos escenarios que van desde la degeneración de las afinidades que lleve al reconocimiento indiferenciado de los BSs alternativos (ambigüedad intrínseca) hasta la existencia de códigos parciales excluyentes (ambigüedad extrínseca). Dado que en la mayoría de los casos no existen datos experimentales para avalar uno u otro escenario, el código incluye, en caso de ambigüedad, la comparación de las frecuencias entre las alternativas palindrómicas y sus mixturas. En general, el código presenta un acusado grado de flexibilidad en la posición AA-15, aunque el residuo AA-16 por sí solo no determina unívocamente la secuencia de nucleótidos contactada. Por último, a partir de la lectura en el código de una de estas degeneraciones ($Y_{15}Q_{16}$ / $H_{15}Q_{16}$) hemos documentado BSs naturales equivalentes a los operadores sintéticos de LacI, contribuyendo así a despejar las dudas que se habían arrojado en cuanto al carácter más o menos artificioso con el que LacI se une a estos operadores.

En conjunto, los resultados conducentes al código son coherentes con los modelos estructurales [30–34] y con los estudios pioneros sobre mutantes de LacI [35,36]. La vigencia de estos modelos sobre buena parte de la familia HTH-LacI refuerza nuestra hipótesis inicial de que el contexto estructural sería lo suficientemente estable a lo largo de la filogenia como para permitir la propia existencia del código.

Epílogo

Esta tesis se inscribe en la pujante disciplina de la Biología de Sistemas, la cual parece destinada a escribirse con idiomas científicos muy variopintos. Como se ha podido comprobar a lo largo de estas páginas, aquí cabe un poco de todo, sean ya técnicas tomadas de la genética clásica o medidas de alta resolución de poblaciones bacterianas; modelos matemáticos en ecuaciones diferenciales o tratamientos de tipo estadístico de las redes de transcripción; o, finalmente, herramientas de genómica comparada. En cierta medida, el abordaje de un mismo problema desde perspectivas y escalas de organización biológica tan distintas hacen que el Biólogo de Sistemas haya de huir en cierta medida de la ultraespecialización que se impone en la ciencia desde hace mucho tiempo, y que, por tanto, lo haga a costa del riesgo no ser un maestro en nada. Aunque son malos tiempos para el humanismo científico, nuestra esperanza radica en que, como acontece en el mismo fenómeno de la vida, el todo que emerja de esta amalgama de disciplinas sea algo más que la suma de las partes.

Bibliografía

[1] Joyce, A. R. and Palsson, B. O. (2006) The model organism as a system: integrating 'omics' data sets. *Nature Rev. Mol. Cell Biol.*, **7**, 198–210.

[2] Boogerd, F. C., Bruggeman, F. J., Hofmeyr, J. S., and Westerhoff, H. V. (2007) *Systems Biology. Philosophical Foundations*. Elsevier, Amsterdam.

[3] Lazebnik, Y. (2002) Can a biologist fix a radio? Or, what I learned while studying apoptosis. *Cancer Cell*, **2**, 179–182.

[4] Guantes, R. and Poyatos, J. F. (2008) Multistable Decision Switches for Flexible Control of Epigenetic Differentiation. *PLoS Computational Biol.*, **4**, e1000235.

[5] Hartwell, L. H., Hopfield, J. J., Leibler, S., and Murray, A. W. (1999) From molecular to modular cell biology. *Nature*, **402**, C47–C52.

[6] Westerhoff, H. V. and Palsson, B. O. (2004) The evolution of molecular biology into systems biology. *Nature Biotechnology*, **22**, 1249–1252.

[7] Savageau, M. A. (1974) Comparison of classical and autogenous systems of regulation in inducible operons. *Nature*, **252**, 546–549.

[8] Wall, M. E., Hlavacek, W. S., and Savageau, M. A. (2004) Design of gene circuits: lessons from bacteria. *Nat Rev Genet*, **5**, 34–42.

[9] Shen-Orr, S. S., Milo, R., Mangan, S., and Alon, U. (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nature Genetics*, **31**, 64–68.

[10] Alon, U. (2007) Network motifs: theory and experimental approaches. *Nat. Rev. Genet.*, **8**, 450–461.

[11] Conant, G. C. and Wagner, A. (2003) Convergent evolution of gene circuits. *Nature Genetics*, **34**, 264–266.

[12] Alon, U. (2003) Biological networks: The tinkerer as an engineer. *Science*, **301**, 1866–1867.

[13] Mazurie, A., Bottani, S., and Vergassola, M. (2005) An evolutionary and functional assessment of regulatory network motifs. *Genome Biol.*, **6**.

[14] Ingram, P. J., Stumpf, M. P. H., and Stark, J. (2006) Network motifs: structure does not determine function. *BMC Genomics*, **7**.

[15] Lynch, M. (2007) The evolution of genetic networks by non-adaptive processes. *Nat. Rev. Genet.*, **8**, 803–813.

[16] Knabe, J. F., Nehaniv, C. L., and Schilstra, M. J. (2008) Do motifs reflect evolved function?-No convergent evolution of genetic regulatory network subgraph topologies. *Biosystems*, **94**, 68–74.

[17] Kashtan, N., Itzkovitz, S., Milo, R., and Alon, U. (2004) Topological generalizations of network motifs. *Phys. Rev. E*, **70**, 031909.

[18] Dobrin, R., Beg, Q. K., Barabási, A. L., and Oltvai, Z. N. (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. *BMC Bioinformatics*, **5**, 10.

[19] Balázsi, G., Barabási, A. L., and Oltvai, Z. N. (2005) Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. *Proc. National Acad. Sciences U S A*, **102**, 7841–7846.

[20] Camas, F. M., Blázquez, J., and Poyatos, J. F. (2006) Autogenous and non-autogenous control of response in a genetic network. *Proc. National Acad. Sciences U S A*, **103**, 12718–12723.

[21] Camas, F. M. and Poyatos, J. F. (2008) What Determines the assembly of transcriptional network motifs in *Escherichia coli*? *PLoS ONE*, **3**, e3657.

[22] Mangan, S., Zaslaver, A., and Alon, U. (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. *J. Mol. Biol.*, **334**, 197–204.

[23] Mangan, S., Itzkovitz, S., Zaslaver, A., and Alon, U. (2006) The incoherent feed-forward loop accelerates the response-time of the *gal* system of *Escherichia coli*. *J. Mol. Biol.*, **356**, 1073–1081.

[24] Kaplan, S., Bren, A., Dekel, E., and Alon, U. (2008) The incoherent feed-forward loop can generate non-monotonic input functions for genes. *Mol. Systems Biol.*, **4**, 203.

[25] Warren, P. B. and ten Wolde, P. R. (2004) Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. *J. Mol. Biol.*, **342**, 1379–1390.

[26] Korbel, J. O., Jensen, L. J., von Mering, C., and Bork, P. (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature Biotechnology*, **22**, 911–917.

[27] Hershberg, R., Yeger-Lotem, E., and Margalit, H. (2005) Chromosomal organization is shaped by the transcription regulatory network. *Trends In Genetics*, **21**, 138–142.

[28] Kolesov, G., Wunderlich, Z., Laikova, O. N., Gelfand, M. S., and Mirny, L. A. (2007) How gene order is influenced by the biophysics of transcription regulation. *Proc. National Acad. Sciences U S A*, **104**, 13948–13953.

[29] Wasserman, W. W. and Sandelin, A. (2004) Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.*, **5**, 276–287.

[30] Schumacher, M. A., Choi, K. Y., Zalkin, H., and Brennan, R. G. (1994) Crystal-structure of LacI member, PurR, bound to DNA: minor-groove binding by α-helices. *Science*, **266**, 763–770.

[31] Glasfeld, A., Koehler, A. N., Schumacher, M. A., and Brennan, R. G. (1999) The role of lysine 55 in determining the specificity of the purine repressor for its operators through minor groove interactions. *J. Mol. Biol.*, **291**, 347–361.

[32] Kalodimos, C. G., Boelens, R., and Kaptein, R. (2004) Toward an integrated model of protein-DNA recognition as inferred from NMR studies on the Lac repressor system. *Chem. Rev.*, **104**, 3567–3586.

[33] Schumacher, M. A., Allen, G. S., Diel, M., Seidel, G., Hillen, W., and Brennan, R. G. (2004) Structural basis for allosteric control of the transcription regulator CcpA by the phosphoprotein HPr-Ser46-P. *Cell*, **118**, 731–741.

[34] Salinas, R. K., Folkers, G. E., Bonvin, A. M. J. J., Das, D., Boelen, R., and Kaptein, R. (2005) Altered specificity in DNA binding by the lac repressor: A mutant lac headpiece that mimics the gal repressor. *ChemBioChem*, **6**, 1628–1637.

[35] Lehming, N., Sartorius, J., Niemöller, M., Geneger, G., von Wilcken-Bergmann, B., and Müller-Hill, B. (1987) The interaction of the recognition helix of lac repressor with *lac* operator. *EMBO J.*, **6**, 3145–3153.

[36] Sartorius, J., Lehming, N., Kisters, B., von Wilcken-Bergmann, B., and Müller-Hill, B. (1989) *lac* repressor mutants with double or triple exchanges in the recognition helix bind specifically to lac operator variants with multiple exchanges. *EMBO J.*, **8**, 1265–1270.

[37] Lewis, M. (2005) The *lac* repressor. *Comptes Rendus Biologies*, **328**, 521–548.

[38] Pabo, C. O. and Nekludova, L. (2000) Geometric analysis and comparison of protein-DNA interfaces: Why is there no simple code for recognition? *J. Mol. Biol.*, **301**, 597–624.

[39] Hall, B. M., LeFevre, K. R., and Cordes, M. H. J. (2005) Sequence correlations between Cro recognition helices and cognate O-R consensus half-sites suggest conserved rules of protein-DNA recognition. *J. Mol. Biol.*, **350**, 667–681.

[40] Zhou, K. and Doyle, J. (1998) *Essentials of Robust Control*. Prentice Hall, New Jersey.

[41] Thieffry, D., Huerta, A. M., Perez-Rueda, E., and Collado-Vides, J. (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *Bioessays*, **20**, 433–440.

[42] Lagomarsino, M. C., Jona, P., Bassetti, B., and Isambert, H. (2007) Hierarchy and feedback in the evolution of the *Escherichia coli* transcription network. *Proc. National Acad. Sciences U S A*, **104**, 5516–5520.

[43] Schwartz, M. (1959) *Information Transmission, Modulation and Noise*. McGraw-Hill, New York.

[44] Ogata, K. (1970) *Modern Control Engineering*. Prentice-Hall, New Jersey.

[45] Brogan, W. L. (1974) *Modern Control Theory*. Prentice-Hall, New Jersey.

[46] Keener, J. and Sneyd, J. (1998) *Mathematical Physiology*. Springer-Verlag, New York.

[47] Guantes, R. and Poyatos, J. F. (2006) Dynamical principles of two-component genetic oscillators. *PLoS Computational Biol.*, **2**, 188–197.

[48] Guido, N. J., Wang, X., Adalsteinsson, D., McMillen, D., Hasty, J., Cantor, C. R., Elston, T. C., and Collins, J. J. (2006) A bottom-up approach to gene regulation. *Nature*, **439**, 856–860.

[49] Suel, G. M., Kulkarni, R. P., Dworkin, J., Garcia-Ojalvo, J., and Elowitz, M. B. (2007) Tunability and noise dependence in differentiation dynamics. *Science*, **315**, 1716–1719.

[50] van Kampen, N. G. (1992) *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam.

[51] Selinger, D. W., Saxena, R. M., Cheung, K. J., Church, G. M., and Rosenow, C. (2003) Global RNA half-life analysis in *Escherichia coli* reveals positional patterns of transcript degradation. *Genome Res.*, **13**, 216–223.

[52] Swain, P. S., Elowitz, M. B., and Siggia, E. D. (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. National Acad. Sciences U S A*, **99**, 12795–12800.

[53] Rosenfeld, N., Elowitz, M. B., and Alon, U. (2002) Negative autoregulation speeds the response times of transcription networks. *J. Mol. Biol.*, **323**, 785–793.

[54] Simpson, M. L., Cox, C. D., and Sayler, G. S. (2003) Frequency domain analysis of noise in autoregulated gene circuits. *Proc. National Acad. Sciences U S A*, **100**, 4551–4556.

[55] Thattai, M. and van Oudenaarden, A. (2001) Intrinsic noise in gene regulatory networks. *Proc. National Acad. Sciences U S A*, **98**, 8614–8619.

[56] Ozbudak, E. M., Thattai, M., Kurtser, I., Grossman, A. D., and van Oudenaarden, A. (2002) Regulation of noise in the expression of a single gene. *Nature Genetics*, **31**, 69–73.

[57] Coquillat, F. (1997) *Cálculo Integral*. Tébar Flores, Madrid.

[58] Becskei, A. and Serrano, L. (2000) Engineering stability in gene networks by autoregulation. *Nature*, **405**, 590–593.

[59] Friedberg, E. C., Walker, G. C., Siede, W., Wood, R. D., and Schultz, R. A. (2006) *DNA Repair and Mutagenesis*. ASM, Washington DC.

[60] Aksenov, S. V., Krasavin, E. A., and Litvin, A. A. (1997) Mathematical model of the SOS response regulation of an excision repair deficient mutant of *Escherichia coli* after ultraviolet light irradiation. *J. Theoretical Biol.*, **186**, 251–260.

[61] Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002) Network motifs: Simple building blocks of complex networks. *Science*, **298**, 824–827.

[62] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004) Superfamilies of evolved and designed networks. *Science*, **303**, 1538–1542.

[63] Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U., and Margalit, H. (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A*, **101**, 5934–5939.

[64] Artzy-Randrup, Y., Fleishman, S. J., Ben-Tal, N., and Stone, L. (2004) Comment on "Network motifs: Simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science*, **305**, 1107c.

[65] Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., and Alon, U. (2004) Response to comment on "Network motifs: Simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science*, **305**, 1107d.

[66] Lee, T. I., et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, **298**, 799–804.

[67] Boyer, L. A., et al. (2005) Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell*, **122**, 947–956.

[68] Iranfar, N., Fuller, D., and Loomis, W. F. (2006) Transcriptional regulation of post-aggregation genes in Dictyostelium by a feed-forward loop involving GBF and LagC. *Developmental Biol.*, **290**, 460–469.

[69] Alon, U. (2007) *An Introduction to System Biology*. Chapmann and Hall/CRC, London.

[70] Isalan, M., Lemerle, C., Michalodimitrakis, K., Horn, C., Beltrao, P., Raineri, E., Garriga-Canut, M., and Serrano, L. (2008) Evolvability and hierarchy in rewired bacterial gene networks. *Nature*, **452**, 840–U2.

[71] Ronen, M., Rosenberg, R., Shraiman, B. I., and Alon, U. (2002) Assigning numbers to the arrows: parameterizing a gene regulation network by using accurate expression kinetics. *Proc Natl Acad Sci U S A*, **99**, 10555–10560.

[72] Strogatz, S. H. (2000) *Nonlinear Dynamics and Chaos: With Applications in Physics, Biology, Chemistry and Engineering*. Perseus Publishing, Massachusetts.

[73] Acar, M., Becskei, A., and van Oudenaarden, A. (2005) Enhancement of cellular memory by reducing stochastic transitions. *Nature*, **435**, 228–232.

[74] Becskei, A. and Serrano, L. (2000) Engineering stability in gene networks by autoregulation. *Nature*, **405**, 590–593.

[75] Sassanfar, M. and Roberts, J. W. (1990) Nature of the SOS-inducing signal in *Escherichia coli*. The involvement of DNA replication. *J Mol Biol*, **212**, 79–96.

[76] Opperman, T., Murli, S., Smith, B. T., and Walker, G. C. (1999) A model for a *umuDC*-dependent prokaryotic DNA damage checkpoint. *Proc Natl Acad Sci U S A*, **96**, 9218–9223.

[77] Banzhaf, W. and Kuo, P. D. (2004) Network Motifs in Natural and Artificial Transcriptional Regulatory Networks. *J. Biol. Phys. Chem.*, **4**, 85–92.

[78] Cordero, O. X. and Hogeweg, P. (2006) Feed-forward loop circuits as a side effect of genome evolution. *Mol. Biol. Evolution*, **23**, 1931–1936.

[79] Ward, J. J. and Thornton, J. M. (2007) Evolutionary models for formation of network motifs and modularity in the *Saccharomyces* transcription factor network. *PLoS Computational Biol.*, **3**, 1993–2002.

[80] Babu, M. M., Teichmann, S. A., and Aravind, L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J. Mol. Biol.*, **358**, 614–633.

[81] Dekel, E. and Alon, U. (2005) Optimality and evolutionary tuning of the expression level of a protein. *Nature*, **436**, 588–592.

[82] Parter, M., Kashtan, N., and Alon, U. (2007) Environmental variability and modularity of bacterial metabolic networks. *BMC Evolutionary Biol.*, **7**, 169.

[83] Price, M. N., Dehal, P. S., and Arkin, A. P. (2008) Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*. *Genome Biol.*, **9**, R4.

[84] Salgado, H., et al. (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res.*, **34**, D394–D397.

[85] Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Royal Statistical Soc. Series B-methodological*, **57**, 289–300.

[86] Zaslaver, A., Mayo, A. E., Rosenberg, R., Bashkin, P., Sberro, H., Tsalyuk, M., Surette, M. G., and Alon, U. (2004) Just-in-time transcription program in metabolic pathways. *Nature Genetics*, **36**, 486–491.

[87] Ma, H. W., Buer, J., and Zeng, A. P. (2004) Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. *BMC Bioinformatics*, **5**, 199.

[88] Yu, H. Y. and Gerstein, M. (2006) Genomic analysis of the hierarchical structure of regulatory networks. *Proc. National Acad. Sciences U S A*, **103**, 14724–14731.

[89] Keseler, I. M., et al. (2009) EcoCyc: A comprehensive view of *Escherichia coli* biology. *Nucleic Acids Res.*, **37**, D464–D470.

[90] Lawrence, J. G. and Roth, J. R. (1996) Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics*, **143**, 1843–1860.

[91] Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M., and Teichmann, S. A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr. Opin. Struct. Biol.*, **14**, 283–291.

[92] Teichmann, S. A. and Babu, M. M. (2004) Gene regulatory network growth by duplication. *Nature Genetics*, **36**, 492–496.

[93] Deutscher, J. (2008) The mechanisms of carbon catabolite repression in bacteria. *Curr. Opin. Microbiol.*, **11**, 87–93.

[94] Kalir, S., Mangan, S., and Alon, U. (2005) A coherent feed-forward loop with a SUM input function prolongs flagella expression in *Escherichia coli*. *Mol. Systems Biol.*, **1**, 2005.0006.

[95] Lewis, M., Chang, G., Horton, N. C., Kercher, M. A., Pace, H. C., Schumacher, M. A., Brennan, R. G., and Lu, P. Z. (1996) Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. *Science*, **271**, 1247–1254.

[96] Ureta-Vidal, A., Ettwiller, L., and Birney, E. (2003) Comparative genomics: Genome-wide analysis in metazoan eukaryotes. *Nat. Rev. Genet.*, **4**, 251–262.

[97] Warren, P. B. and ten Wolde, P. R. (2004) Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. *J. Mol. Biol.*, **342**, 1379–1390.

[98] Alm, E. J., Huang, K. H., Price, M. N., Koche, R. P., Keller, K., Dubchak, I. L., and Arkin, A. P. (2005) The MicrobesOnline web site for comparative genomics. *Genome Res.*, **15**, 1015–1022.

[99] Choi, K. Y. and Zalkin, H. (1994) Role of the purine repressor hinge sequence in repressor function. *J. Bacteriol.*, **176**, 1767–1772.

[100] Bell, C. E. and Lewis, M. (2001) The Lac repressor: a second generation of structural and functional studies. *Curr. Opin. Struct. Biol.*, **11**, 19–25.

[101] Kazakov, A. E., Cipriano, M. J., Novichkov, P. S., Minovitsky, S., Vinogradov, D. V., Arkin, A., Mironov, A. A., Gelfand, M. S., and Dubchak, I. (2007) RegTransBase - a database of regulatory sequences and interactions in a wide range of prokaryotic genomes. *Nucleic Acids Res.*, **35**, D407–D412.

[102] Francke, C., Kerkhoven, R., Wels, M., and Siezen, R. J. (2008) A generic approach to identify Transcription Factor-specific operator motifs; Inferences for LacI-family mediated regulation in *Lactobacillus plantarum* WCFS1. *BMC Genomics*, **9**.

[103] Stormo, G. D. (2000) DNA binding sites: representation and discovery. *Bioinformatics*, **16**, 16–23.

[104] Betz, J. L., Sasmor, H. M., Buck, F., Insley, M. Y., and Caruthers, M. H. (1986) Base substitution mutants of the *lac* operator - invivo and invitro affinities for *lac* repressor. *Gene*, **50**, 123–132.

[105] Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

[106] Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P., and Bork, P. (2000) SMART: a web-based tool for the study of genetically mobile domains. *Nucleic Acids Research*, **28**, 231–234.

[107] Thompson, W., Rouchka, E. C., and Lawrence, C. E. (2003) Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res.*, **31**, 3580–3585.

[108] Noreen, E. W. (1989) *Computer-Intensive Methods for Testing Hypotheses: An Introduction*. Wiley-Interscience.

Apéndice A

Transmisión de señal y ruido en sistemas lineales autorregulados negativamente

En este Apéndice vamos a presentar de manera sucinta algunos conceptos y resultados fundamentales que atañen a los sistemas lineales con autorregulación negativa. La descripción matemática es general y su aplicación a un sistema biológico se aborda en el Capítulo 1. De hecho, lo que aquí tratamos supone sólo un pequeño extracto del vasto corpus teórico de los sistemas de control basados en la realimentación, corpus que se desarrolló en los terrenos de la ingeniería industrial y de las telecomunicaciones. Por ello, los manuales más completos al respecto han de buscarse entre los de estas disciplinas [43–45].

Empecemos, pues, por la siguiente ecuación, que describe la dinámica de un sistema cuya respuesta $x(t)$ realimenta negativamente la entrada de éste o, dicho de otro modo, la dinámica de un sistema con autorregulación negativa,

$$\frac{\mathrm{d}x}{\mathrm{d}t} = \mathrm{B}\,u(t) - \mathrm{A}\,x(t) \quad \text{con} \quad A > 0. \tag{A.1}$$

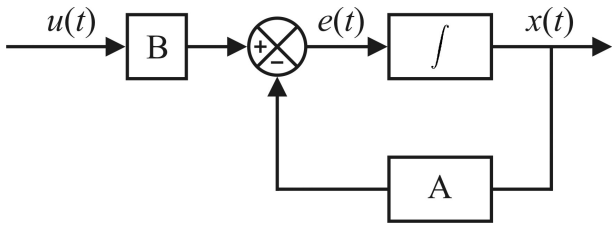


Figura A.1: *Diagrama de bloques del sistema lineal (A.1).*

La Figura A.1 muestra el diagrama de bloques para esta ecuación lineal[1], donde parámetros y variables se reinterpretan en base a conceptos que atañen al flujo de las señales. Así, $u(t)$ es la entrada de referencia; $x(t)$, la variable controlada o salida; A, la función de transferencia en el trayecto de realimentación; B, la función de transferencia en el trayecto de entrada; $e(t) = \mathrm{B}\,u(t) - \mathrm{A}\,x(t)$, la señal de error. Finalmente, el bloque entre $e(t)$ y $x(t)$ es función de transferencia en bucle abierto. En este caso el signo de integración indica que la señal de error se integra en el tiempo para producir $x(t)$.

A.1. Transmisión de señal

Vamos a comenzar el estudio del comportamiento de la respuesta $x(t)$ del sistema aplicando a la entrada del circuito una señal escalón de altura u tal que

$$u(t) = \begin{cases} 0 & t < 0 \\ u & t \geq 0. \end{cases}$$

En este caso, la solución analítica de la ecuación (A.1) es

$$x(t) = (1 - e^{-t/\tau_r})\,\frac{\mathrm{B}}{\mathrm{A}}\,u,$$

donde $\tau_r \equiv 1/\mathrm{A}$ define el tiempo de respuesta característico del sistema. En particular, la solución estacionaria vale

$$x_* = \frac{\mathrm{B}}{\mathrm{A}}\,u$$

Así que cuanto mayor es el valor de la autorregulación A menor es τ_R (se alcanza antes el estacionario) y menor es el valor absoluto de la *ganancia*[2] en el nuevo estado estacionario $|x_*/u| = |\mathrm{B}|/\mathrm{A}$ (Figura A.2). La existencia del tiempo característico τ_R resulta determinante ante señales que están continuamente variando en el tiempo, puesto que el sistema no va a poder responder ante cambios que se produzcan de un modo arbitrariamente rápido, es decir, que su respuesta $x(t)$ no será capaz de reflejar los cambios de frecuencia críticamente alta que se produzcan en la entrada $u(t)$. No obstante, el rango de las frecuencias ante las que el circuito será sensible –lo que se denomina el *ancho de banda* del circuito– será mayor cuanto más fuerte sea la autorregulación. Para ver esto de manera

[1] En el contexto de la ingeniería de control, el diagrama de bloques es una herramienta muy utilizada por su virtud de trasladar las ecuaciones diferenciales a un lenguaje visual que permite seguir el flujo de las señales [44].

[2] La función de transferencia B no tiene por que ser necesariamente positiva. Para $\mathrm{B} < 0$ se produce simplemente una inversión en el signo de la respuesta respecto al de la señal de entrada.

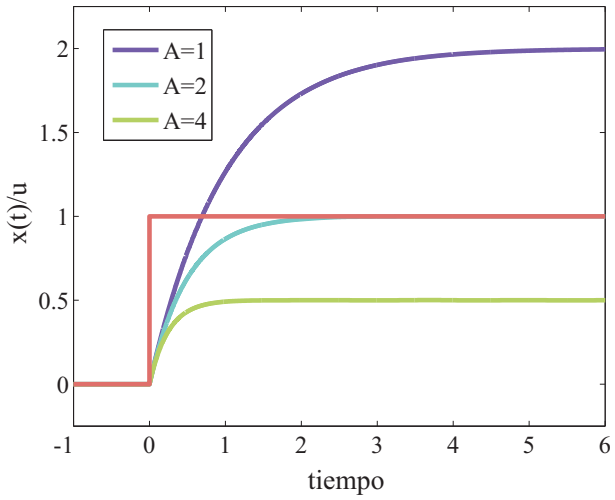


Figura A.2: *Respuesta del sistema a una señal de entrada tipo escalón (en rojo) para distintos valores de la constante de autorregulación* A*. El valor de* B *es siempre igual a 2. Las curvas están normalizada por la altura del escalón u.*

cuantitativa estudiemos el comportamiento del sistema ante una onda armónica $u(t) = u\, e^{i2\pi f\, t}$, lo cual supone una generalización esencial en el estudio de las propiedades de transmisión del sistema[3].

Para movernos en el dominio de las frecuencias vamos a realizar la transformación de Fourier de la ecuación (A.1):

$$i2\pi f X(f) = \mathrm{B}\, U(f) - \mathrm{A}\, X(f), \tag{A.2}$$

siendo $X(f)$ e $U(f)$ las variables transformadas de $x(t)$ e $y(t)$, respectivamente. La forma lineal de (A.2) permite directamente relacionar las variables de entrada y salida en lo que se define como la *función de transferencia en bucle cerrado*,

$$H(f) \equiv \frac{X(f)}{U(f)} = \frac{H_o}{1 + i\,\frac{f}{f_o}} \quad \text{con} \quad H_o \equiv H(0) = \frac{\mathrm{B}}{\mathrm{A}} \quad \text{y} \quad f_o = \frac{\mathrm{A}}{2\pi}. \tag{A.3}$$

La proporción (o ganancia) entre las amplitudes de las ondas armónicas de entrada y salida se obtiene tomando el módulo de H,

$$|H(f)| = |H_o| \frac{1}{\sqrt{1 + \left(\frac{f}{f_o}\right)^2}}. \tag{A.4}$$

[3]Puesto que toda señal $n(t)$ de duración T puede ser descompuesta en su serie de Fourier:

$$n(t) = \frac{1}{T} \sum_{n=-\infty}^{\infty} c_n e^{i2\pi f_n t} \quad con \quad f_n = \frac{n}{T} \quad y \quad c_n = \int_{-T/2}^{T/2} n(t) e^{i2\pi f_n t} \mathrm{d}t.$$

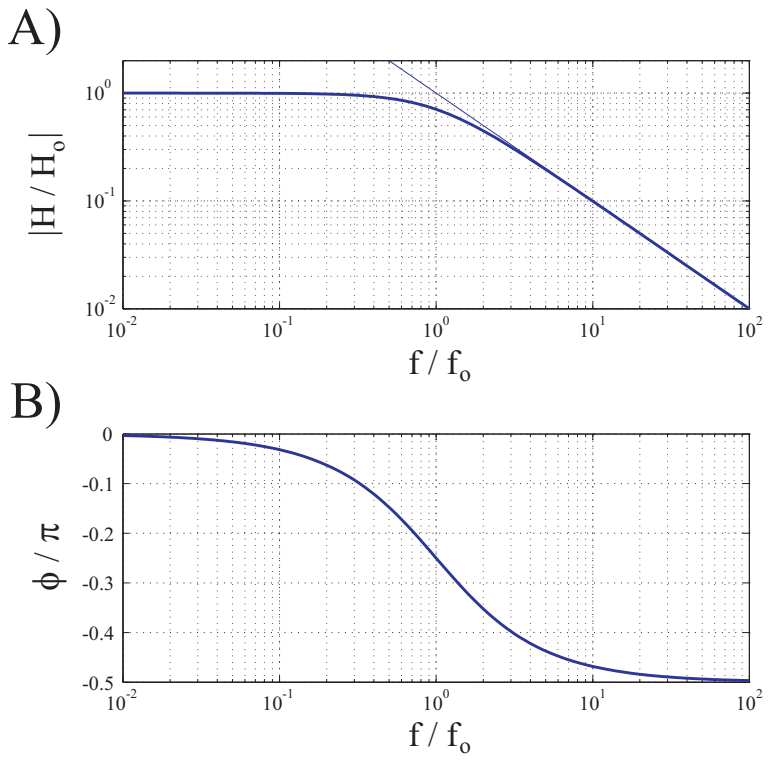


Figura A.3: *A) Ganancia en función de la frecuencia. Se muestra también la asíntota para el régimen $f \gg f_o$. B) Fase entre las ondas de entrada y salida. Las frecuencias están normalizadas por la frecuencia característica f_o. El conjunto de estas dos gráficas, que caracteriza las propiedades de filtrado del sistema, se conoce como diagrama de Bode.*

Esta expresión aparece representada en la Figura A.3.A. Nótese que a bajas frecuencias ($f \ll f_o$) se recupera la solución estacionaria para la señal escalón, $|H_o| = |\mathrm{B}|/\mathrm{A}$. En cambio, para altas frecuencias ($\frac{f}{f_o} \gg 1$) se tiene que $|H(f)/H_o| \sim 1/f$, lo cual da lugar en la representación logarítmica a un comportamiento asintóticamente lineal con pendiente igual a -1. Por otro lado, la relación entre las fases de las ondas de entrada y salida viene dada por

$$\phi(f) = \arg H(f) = -\arctan \frac{f}{f_o},$$

y aparece representada en la Figura A.3.B. Finalmente, la forma de la onda armónica a la salida del circuito es

$$x(t) = u|H(f)|\, e^{i(2\pi f\, t+\phi)}.$$

A.2. Transmisión de ruido

Hasta ahora hemos considerado señales de entrada $u(t)$ de carácter determinista; pero, ¿qué sucede si la señal $u(t)$ es de tipo aleatorio, esto es, un ruido? La naturaleza estocástica de la señal exige entonces tratarla en términos estadísticos, en concreto, en términos de su densidad espectral de potencia (PSD, por *power spectral density*)[4].

Una señal con una PSD constante para toda frecuencia, $G_u(f) = G_u$, es lo que se conoce como *ruido blanco*, e implica que esta señal tendría teóricamente una varianza infinita[5], $\sigma_u^2 = \infty$. No obstante, aunque toda señal real tiene una PSD que decae para frecuencias suficientemente altas, cuando el propio sistema impone una limitación aún más severa al contenido frecuencial de la respuesta $x(t)$, la asunción de un ruido blanco en la señal de entrada es una buena aproximación.

La PSD es una magnitud cuadrática en las unidades de la señal y cuadrática es su transmisión a través del circuito, pues se puede demostrar [43] que el ruido a la salida del circuito tiene una PSD dada por,

$$G_x(f) = |H(f)|^2 G_u(f).$$

Si usamos ahora la forma de la función de transferencia (A.4) y si colocamos un ruido blanco a la entrada del circuito, tenemos que

$$G_x(f) = \frac{H_o^2}{1 + \left(\frac{f}{f_o}\right)^2} G_u,$$

donde ya se observa el efecto de atenuación de la PSD en las frecuencias altas que impone el circuito (Figura A.4.A). Esto se va a traducir en una varianza (ahora

[4]Brevemente, si $R_u(\tau) \equiv \langle u(t)u(t+\tau)\rangle$ es la función de autocorrelación de $u(t)$, la PSD $G_u(f)$ es su transformada de Fourier:

$$R_u(\tau) = \int_{-\infty}^{\infty} G_u(f) e^{-i2\pi f\tau} \mathrm{d}f \qquad G_u(f) = \int_{-\infty}^{\infty} R_u(\tau) e^{i2\pi f\tau} \mathrm{d}\tau.$$

Si el promedio de la señal es nulo $\langle u(t)\rangle = 0$, la varianza de la señal es igual a la función de correlación evaluada en $\tau = 0$, $\sigma_u^2 = R_u(0)$. Por tanto, en ese caso,

$$\sigma_u^2 = \int_{-\infty}^{\infty} G_u(f)\, \mathrm{d}f,$$

expresión en la que queda clara la razón del nombre de densidad espectral de potencia, puesto que $G_u(f)$ es la densidad de probabilidad de la varianza en función de la frecuencia [43].

[5]Es decir, una energía infinita y una total descorrelación entre $u(t)$ y $u(t+\tau)$ por pequeño que sea τ, puesto que

$$R_u(\tau) = \int_{-\infty}^{\infty} G_u e^{-i2\pi f\tau} \mathrm{d}f = G_u \delta(0).$$

sí) finita para la señal de salida[6]:

$$\sigma_x^2 = \int_{-\infty}^{\infty} G_x(f)\,\mathrm{d}f = 2\Delta f\, H_o^2\, G_u \tag{A.5}$$

siendo Δf el *ancho de banda equivalente* del circuito,

$$\Delta f \equiv \int_0^{\infty} \frac{|H(f)|^2}{H_o^2}\,\mathrm{d}f = \int_0^{\infty} \frac{1}{1+\left(\frac{f}{f_o}\right)^2}\,\mathrm{d}f = \frac{\pi}{2}\,f_o = \frac{\mathrm{A}}{4},$$

donde el adjetivo *equivalente* le viene por ser el ancho de banda que daría lugar al mismo valor de la varianza en un filtro *paso de baja* ideal (Figura A.4.B)[7], esto es,

$$|H(f)|^2 = \begin{cases} H_o^2 & \text{si } |f| \in [0, \Delta f] \\ 0 & \text{si } |f| > \Delta f. \end{cases}$$

La imagen del filtro ideal ofrece una interpretación geométrica inmediata de (A.5): aunque el ancho de banda crece linealmente con la autorregulación A, la ganancia a frecuencia cero, que contribuye cuadráticamente al valor del ruido, es inversamente proporcional a la misma –véase (A.3)–, con lo que el área bajo el filtro ideal, $2\,\Delta f\, H_o^2$, es también inversamente proporcional a A. Efectivamente, sustituyendo en (A.5) los valores de H_o y el ancho de banda, obtenemos para la varianza de la señal a la salida del circuito:

$$\sigma_x^2 = \frac{\mathrm{B}^2}{2\mathrm{A}}\,G_u.$$

La ecuación anterior recoge una propiedad esencial de los circuitos autorregulados: cuanto mayor es la autorregulación menor es el ruido a la salida del circuito.

De todos modos, la interpretación en base a filtros ideales tiene también sus limitaciones, pues tomando la imagen de los dos que aparecen en la Figura A.4.B en toda su literalidad podríamos llegar a la conclusión de que cuanto más laxa sea la autorregulación más restringido será el espectro de frecuencias a la salida del circuito. Es decir, que el circuito más fuertemente autorregulado de esta Figura dejaría pasar un tramo de frecuencias altas inaccesibles al primero. No obstante, cuando tenemos en cuenta la forma real de los filtros de la Figura A.4.A, vemos que una mayor autorregulación implica en realidad menores amplitudes a lo largo

[6] Y un tiempo de autocorrelación del orden de $1/\mathrm{A} = \tau_R$,

$$R_x(\tau) = \int_{-\infty}^{\infty} G_x(f) e^{-i2\pi f\tau}\mathrm{d}f = \sigma_x^2 e^{-A|\tau|}.$$

[7] Obsérvese que el ancho de banda se define sólo en las frecuencias positivas, de ahí el factor 2 que aparece en (A.5).

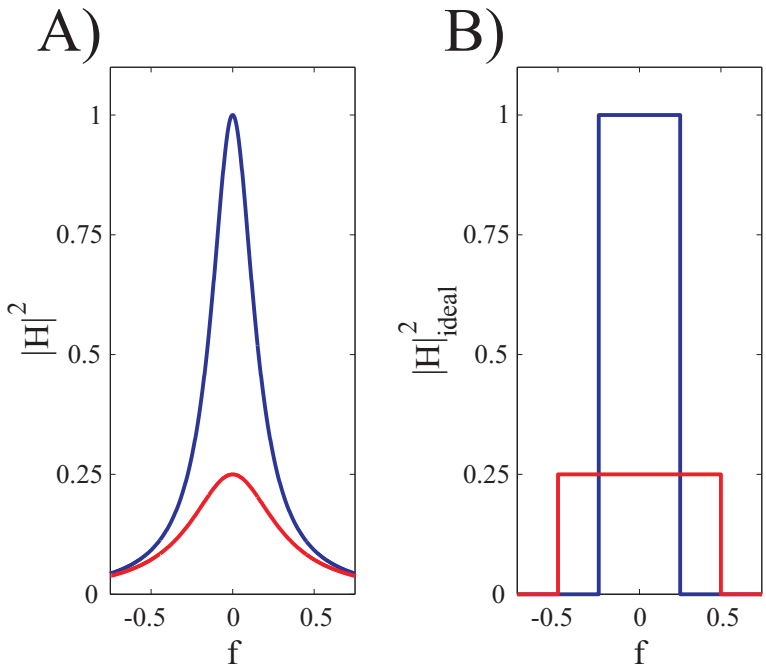


Figura A.4: *La función de transferencia cuadrática* $|H(f)|^2$ *relaciona las PSDs de entrada y de salida, filtrando la primera. A)* $|H(f)|^2$ *para* A=1 *(línea azul) y* A=2 *(en rojo). En ambos casos* B=1. *Los anchos de banda respectivos son* Δf=0.25 *y* Δf=0.5. *B) Equivalentes* ideales *de los dos filtros de A).*

de todo el espectro. De manera más general, esto se comprueba fácilmente si tomamos dos circuitos con autorregulaciones A_1 y A_2 tal que $A_1 < A_2$ y que tengan un mismo valor de parámetro de entrada ($B_1 = B_2 = B$). La relación entre los cuadrados de las funciones de transferencia sería entonces

$$\frac{|H_1(f)|^2}{|H_2(f)|^2} = \frac{A_2^2 + (2\pi f)^2}{A_1^2 + (2\pi f)^2},$$

expresión que es siempre mayor que la unidad para todas las frecuencias por ser $A_1 < A_2$ y que alcanza su máximo valor $(A_2/A_1)^2$ para $f = 0$ (Figura A.4.A).

Apéndice B

Autogenous and nonautogenous control of response in a genetic network

(*Artículo*)

Autogenous and nonautogenous control of response in a genetic network

Francisco M. Camas[†‡], Jesús Blázquez[†§], and Juan F. Poyatos[‡¶]

[†]Spanish National Biotechnology Centre (CNB)–Consejo Superior de Investigaciones Científicas (CSIC), 28049 Madrid, Spain; and
[‡]Spanish National Cancer Centre (CNIO), 28029 Madrid, Spain

Edited by Charles R. Cantor, Sequenom, Inc., San Diego, CA, and approved July 6, 2006 (received for review March 15, 2006)

Feedback-based control methods determine the behavior of cellular systems, an example being autogenous control, the regulation of production of a protein by itself. This control strategy was theoretically shown to be superior to an equivalent but nonautogenously regulated system when based on a repressor. Although some of its advantages were later confirmed with isolated synthetic circuits, the superiority of autogenous control in natural networks remains untested. Here, we use the SOS DNA repair system of *Escherichia coli*, where autogenous control is part of a single-input module, as a valid model to evaluate the functional advantages and biological implications of this mechanism. We redesign the control of its master regulator, the protein LexA, so that it becomes nonautogenously controlled. We compare both systems by combining high-resolution expression measurements with mathematical modeling. We show that the stronger stability associated with the autogenous regulation prevents false triggering of the response due to transient fluctuations in the inducing signal and that this control also reduces the system recovery time at low DNA damage. Likewise, autoregulation produces responses proportional to the damage signal level. In contrast, bacteria with LexA constitutively expressed induce maximal action even for very low damage levels. This excess in response comes at a cost, because it reduces comparatively the growth rate of these cells. Our results suggest that autogenous control evolved as a strategy to optimally respond to multiple levels of input signal minimizing the costs of the response and highlights reasons why master regulators of single-input modules are mostly autorepressed.

autogenous regulation | design principles | feedback control | synthetic biology | systems biology

The regulatory complexity of cellular systems is attained by the application of different feedback control strategies. Such mechanisms, even when implemented by simple genetic circuits, are often associated with complex dynamical behaviors, whose complete characterization is necessary to better comprehend fundamental cellular actions. Examples of these schemes are increasingly being discovered by the following two complementary approaches. First, because of the extensive characterization of the molecular constituents of some systems over the years, e.g., the mitogen-activated protein kinase signaling cascade (1) or the network of mediators in the maturation of *Xenopus* oocytes (2), it is now possible to analyze their control-level properties. Alternatively, studies of large regulatory networks assembled by using genomic data were able to extract recurrent control architectures used in such networks, e.g., control motifs found in the transcriptional network of *Escherichia coli* (3) or *Saccharomyces cerevisiae* (4). As a whole, these series of findings are putting forward design principles that resume our understanding of the relationship between structure and dynamics of control mechanisms (5–14), which can be applicable to a wide variety of biological contexts. For instance, the union of various feed-forward loops might be a common feature of programs of cellular differentiation (9, 15). More recently, a combination of fast and slow positive feedback loops were shown to enhance the reliability of cell decisions (16).

The functioning of these control schemes in nonbiological scenarios is commonly associated with optimized system performance (17), optimization that in biological systems has been argued to be a consequence of evolution (13, 18, 19). Thus, one could wonder whether the modification of extant control structures in genetic networks would lead to unoptimized regulation of their corresponding biological responses. Among the many feedback-based methods found in cellular networks, autogenous regulation is probably one of the simplest structures to use for analyzing this issue. Indeed, theoretical comparisons proposed that, when based on a repressor, autogenously regulated systems would exhibit several functional advantages with respect to nonautogenously regulated ones, i.e., they should be generally considered more optimized (20, 21). The development of minimal synthetic circuits confirmed some of these aspects experimentally (22, 23). However, these features were studied in isolated synthetic networks, and thus it remains unknown the extent to which natural networks are found in the precise regime where the superiority of autogenous control becomes effective. Here, we characterize the dynamics of these two control strategies, with all other aspects of the system being equal, and suggest the reasons for their selection in a particular natural regulatory network. Our study also sheds light on why the autoregulation of the master transcriptional factor of a single-input module is generally autorepression (3).

To address these issues, we compared the dynamics of the SOS regulatory network in *E. coli* under either type of control. The SOS genetic network activates the response to DNA damage in many bacterial organisms. It is composed of a set of genes under synchronized transcriptional regulation of the LexA protein. This master regulator exhibits autogenous control because it represses its own production. DNA damage caused by different agents, e.g., UV radiation, acts as an inducing signal to the system. Exposure to this radiation causes single-stranded DNA (ssDNA), which activates a second core element of the response: the RecA protein. Activated RecA (RecA*) promotes autocleavage of LexA, originating a gradual increase in transcription of the SOS genes. This increase triggers several mechanisms able to repair the damage and hence diminish the concentration of activated RecA. All of these processes ultimately bring LexA back to its original level.

Results

We studied the control strategy of the SOS master regulator by contrasting the natural system with a synthetic one in which LexA production is nonautogenously regulated. In this latter circuit, it was

Conflict of interest statement: No conflicts declared.

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: PA, promoter activity; IPTG, isopropyl-β-D-thiogalactopyranoside.

[§]To whom correspondence may be addressed at: Microbial Biotechnology Department, Spanish National Biotechnology Centre (CNB), 28049 Madrid, Spain. E-mail: blazquez@cnb.uam.es.

[¶]To whom correspondence may be addressed at: Evolutionary Systems Biology Initiative, Structural and Computational Biology Programme, Spanish National Cancer Centre (CNIO) Melchor Fernández Almagro, 3, 28029 Madrid, Spain. E-mail: jpoyatos@cnio.es.

www.pnas.org/cgi/doi/10.1073/pnas.0602119103

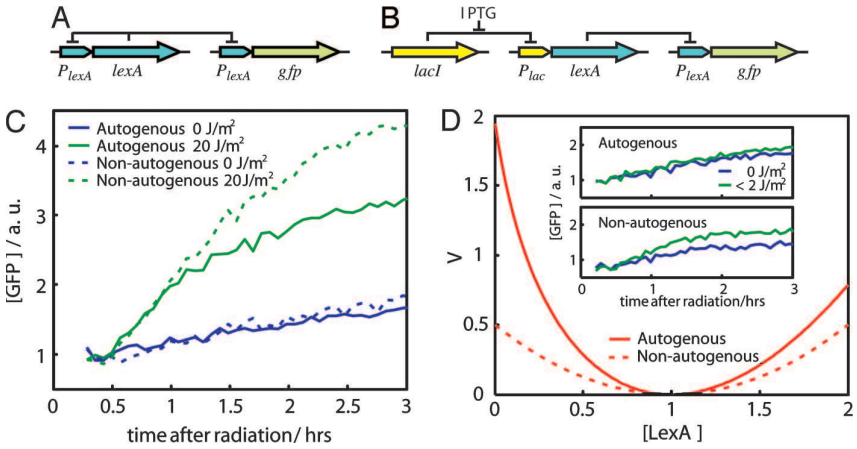


Fig. 1. Autogenous and nonautogenous control of LexA determines SOS response. (*A*) Wild-type LexA regulatory circuit plus GFP reporter. (*B*) Synthetic LexA regulatory circuit plus GFP reporter. (*C*) SOS dynamics (GFP fluorescence) after UV irradiation in cells with (solid line) or without (dashed line) autogenous regulation. GFP levels without radiation for both cases are shown in blue. (*D*) Stability potential, *V* (26, 27), associated with the uninduced LexA state (before UV radiation) for the autogenous (solid line) and nonautogenous (dashed line) circuits. LexA concentrations are in units of uninduced LexA equilibrium such that $x_{eq} = 1$. (*D Inset*) SOS dynamics as in *C* for a very small UV dose (<2 J/m^2) reveals the buffering effect of autogenous control. Reproducibility error of experiments performed on different days was ≈10%.

necessary to reproduce the same level of LexA protein as that found in the natural case so that both systems only differ *a priori* in their control mechanism (Fig. 1 *A* and *B*). To this aim, LexA protein was produced under the control of a lactose promoter tightly regulated by the LacI repressor in a *lexA*(Def) strain. LexA dosage can thus be externally manipulated by induction with isopropyl-β-D-thiogalactopyranoside (IPTG). We obtained cell growth (in terms of optical density; OD) and expression measurements (by using a low-copy reporter plasmid) at high resolution. In this plasmid, GFP is under the control of the *lexA* promoter. We can in this way measure how LexA concentration evolves as a quantity proportional to the rate of GFP accumulation in both circuits [promoter activity (PA); see *Materials and Methods*]. Note that whereas this PA is related to the rate of lexA transcription in both systems, this rate is fixed for the IPTG-induced one. Monitoring of LexA dynamics allowed us to characterize quantitatively the differences in the course of the SOS induction associated with the two control strategies.

Stability and Buffering. We first monitored the response to a fixed dose of UV radiation for both systems. This radiation originates the critical signal (ssDNA) to drive LexA levels from an initially uninduced concentration (before UV) to a lower induced level (after UV) (24). This second concentration represents an equilibrium between degradation, i.e., cleavage, and synthesis (25). As LexA levels diminish, protein production remains fixed in a nonautogenously regulated system, whereas it increases in the presence of feedback control, acting in this latter case as a more effective compensatory force. As a consequence, the balance between production and degradation is achieved at higher LexA levels for the autogenous circuit, and the resulting induced levels are drastically different. This difference is shown in Fig. 1*C*, where we plotted the dynamics of the reporter GFP for both circuits as a function of time after radiation. A lower induced LexA concentration implies a higher GFP level as repression of GFP production is released.

The previous behavior can be partially understood in terms of the gain in stability of the uninduced steady state of the autogenous circuit. Theoretical results have shown how autoregulation enhances the stability of genetic circuits (20). This feature has been experimentally tested with the use of simple synthetic elements (22). We were interested in observing this effect in a natural network. For this process, a small dose of UV radiation can be considered as an external perturbation experienced by the uninduced steady state of LexA. This perturbation is able to move the protein level out of its equilibrium. The stability σ of each system would then determine the strength of this displacement, i.e., the network strength of response. By considering two simple mathematical models of LexA production, with and without autogenous control, the ratio of stabilities $\theta = \sigma_{au}/\sigma_{na}$ can be shown to be proportional to the PAs of both circuits $\theta = 2 - \alpha_{na}/\alpha_{au}$, with α_{na} being the constant PA of the nonautogenous circuit and α_{au} the PA in the absence of repressor of the autogenous one (see *Materials and Methods*). For a strong feedback, $\alpha_{na}/\alpha_{au} \ll 1$, and the ratio θ reaches its maximum (2-fold). We can estimate the value of this ratio by using of the experimental data on the activity of the promoter (see below for details). The experimental value ($\theta \approx 1.8 \pm 0.1$) is close to the theoretical maximum, which indicates that the autoregulation of the LexA system is fully exploiting this effect. In summary, one could interpret that for low-UV doses, the weaker the stability of the system, the stronger the response. We can further visualize this result by representing the time evolution of LexA concentration as the motion of a heavily damped particle in a potential well (26, 27). The stability of the uninduced LexA state is linked to the curvature of the potential (stronger curvatures implying stronger stability; see Fig. 1*D*). Finally, the difference in stability is also reflected in the efficient buffering against fluctuations in the inducing signal by the autoregulatory system. We tested this phenomenon by irradiating cells with minimal UV doses (<2 J/m^2). An appreciable effect of this radiation is only seen for the nonautogenous system (Fig. 1*D Inset*).

Recovery Dynamics. We studied the recovery dynamics of both systems by analyzing their PA. The transition between the induced and uninduced levels of LexA is influenced by both the network intrinsic regulatory dynamics and the repair processes. In this sense, one could envisage two opposite situations. For low damage, activated RecA induces LexA cleavage only for a very short period, because DNA damage is very quickly repaired. The recovery to the uninduced LexA level is mainly determined by the response time of a transcriptional unit whose protein level is out of equilibrium. Both types of control should then differ in their recovery times (23). Alternatively, for larger UV doses, the recovery time is directly related to the repair time. The transition to the uninduced state is mainly driven by the gradual disap-

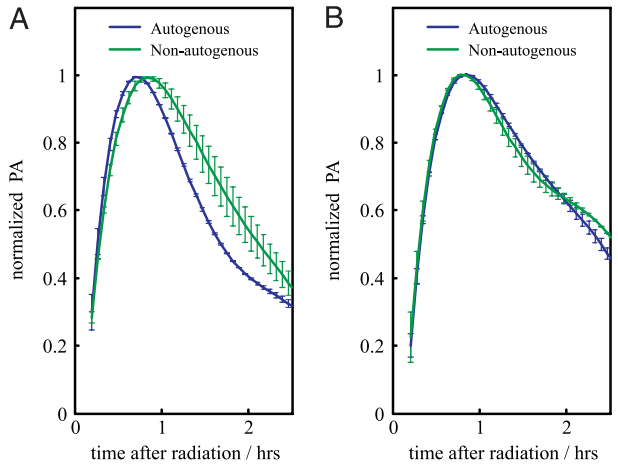


Fig. 2. Response dynamics of the SOS system after UV irradiation in terms of normalized PA. (*A*) PA of the autogenous (blue) and nonautogenous (green) circuits in a regime of low-UV dose (8 J/m^2) clearly differs. (*B*) As in *A* for a regime of high-UV dose (70 J/m^2). In this case, both circuits exhibited very similar recovery times. See text for details.

pearance of the LexA cleavage effect. In this scenario, there should not be a major difference between either type of control.

To validate these hypotheses, we computed the PA of each circuit for a given UV dose by taking the derivative of a polynomial fit of the GFP dynamics (we also applied a regression spline algorithm; see *Materials and Methods*). Maximal PA was different in each system, especially for the low-UV dose (see *Discussion*). To analyze the turn-on (-off) of the response and to measure recovery time, PAs were normalized by their corresponding maximal value in each situation (9). In Fig. 2, we plot the normalized PA for the autogenously and nonautogenously controlled circuits in a situation of low (Fig. 2*A*) and high (Fig. 2*B*) damage. The turn-on transition (uninduced to induced) is very similar for all UV doses and independent of the circuit type. This result is in agreement with previous reports linking this transition to the fast LexA cleavage dynamics (25). However, the turn-off transition (induced to uninduced) determining the recovery time is different. We find that recovery time is smaller for the autogenously regulated circuit at low-UV doses (Fig. 2). We suggest that, in this regime, the repair process is fast enough that the kinetics associated with autogenous control shortens the recovery time of the corresponding circuit. For increasing UV dose, both systems exhibit similar recovery times, which seems to support the initial hypothesis, i.e., that strong DNA damage recovery time is determined by the repair processes.

Range of Response. Exposure of both circuits to a range of UV doses revealed differences in the proportionality of the SOS response to the intensity of the inducing signal (24, 28). To quantify this phenomenon, cells were given different doses of UV radiation. The nonautogenous system always exhibited an overresponse with respect to the autogenous one. We plotted a comparison of the action of both circuits for low and high damage (Fig. 3 *A* and *B*). The relative difference in reaction was revealed to be stronger for low-UV doses.

The strength of the response is associated with the minimum of LexA concentration, which corresponds to a maximum in PA (see *Materials and Methods*). In Fig. 3*C*, we show how this value changes for different UV doses. The ability of the system to exhibit a graded response is clearly lost in the case of nonautogenously regulated LexA. We introduce a simple mathematical model to better understand this behavior (see *Materials and Methods*). The model describes the dynamics of LexA and activated RecA considering a fixed amount of damage; i.e., we did not include any sort of repair mechanism that decreases damage with time. The equilibrium concentrations of this model correspond to a situation briefly after UV radiation, and thus they denote maximal strength of response (minimal LexA). The difference in response between the autogenous and nonautogenous systems can be understood in terms of the geometry of the associated response curves (nullclines; Fig. 3*D*) in the phase space (26). These curves depict the equilibrium concentration of LexA (d[LexA]/dt = 0) and activated RecA (d[RecA*]/dt = 0), respectively. We see how the LexA nullcline decreases more gradually as activated RecA increases for the autogenous circuit. This result implies that for this system the equilibrium points, i.e., the intersection with the activated RecA nullcline, are distributed in a wider range of LexA values with respect to the nonautogenous circuit. In this latter case, these states are also shifted to lower LexA levels. The geometry of the nullclines derived from this simple model highlights why larger differences in response are found at low damage.

Differential Growth Rate. It has been reported that the products of two SOS genes, *umuD* and *umuC*, can slow down the cell cycle in *E. coli*, allowing more time for DNA repair (29). This effect would be particularly important under conditions in which the SOS system is strongly activated. A disproportionate SOS response might thus penalize cell growth, reducing bacterial fitness. We wanted to address the following question: Would the difference in activation between the nonautogenous and autogenous systems be reflected in a differential growth rate under the same damage conditions?

We tested this hypothesis by measuring cell growth in terms of OD. In Fig. 4*A*, we plot the growth of cells with the autogenously regulated LexA for different UV values. We show how the exposure to increasing levels of UV radiation is reflected in a gradual reduction of the growth rate. This very same feature is observed in cells carrying the nonautogenously regulated LexA (Fig. 4*B*). However, the overresponse associated with this type of control, in comparison with an equivalent damage situation with autoregulated LexA, comes at a cost, because the relative growth rate of the population is smaller. This cost is more evident at low-UV doses where the dynamical differences between both circuits are more drastic as discussed above.

Discussion

Biomolecular interactions within cells ultimately decide their physiology. This genetic circuitry resembles in many aspects that found in other nonbiological scenarios, and thus control ideas commonly used in these contexts have been incorporated into the understanding of cellular action (30). These types of studies are contributing to the discovery of a set of feedback-based regulatory strategies in biological systems and to further confirmation of the possible identification of fundamental design principles of cellular control [e.g., robustness (5, 6), noise tolerance (7), programmed temporal order (8, 11), sign-sensitive dynamics (9, 10), ultrasensitivity (12), optimal performance (13), and implementation-dependent dynamics (14)].

To fully understand cellular control mechanisms, it becomes necessary both to characterize their dynamic properties in theoretical terms and to confirm some of these aspects experimentally (21, 31). Although the use of synthetic circuits in isolation is a valid tool for this second goal, and an interesting development by itself from a bioengineering perspective, only by studying these structures as part of natural networks can we completely understand their biological relevance. Here, we characterize one of these strategies, autogenous control, by combining mathematical models and high-resolution expression measurements from living cells. We consider the SOS response network as a valid genetic model for this study

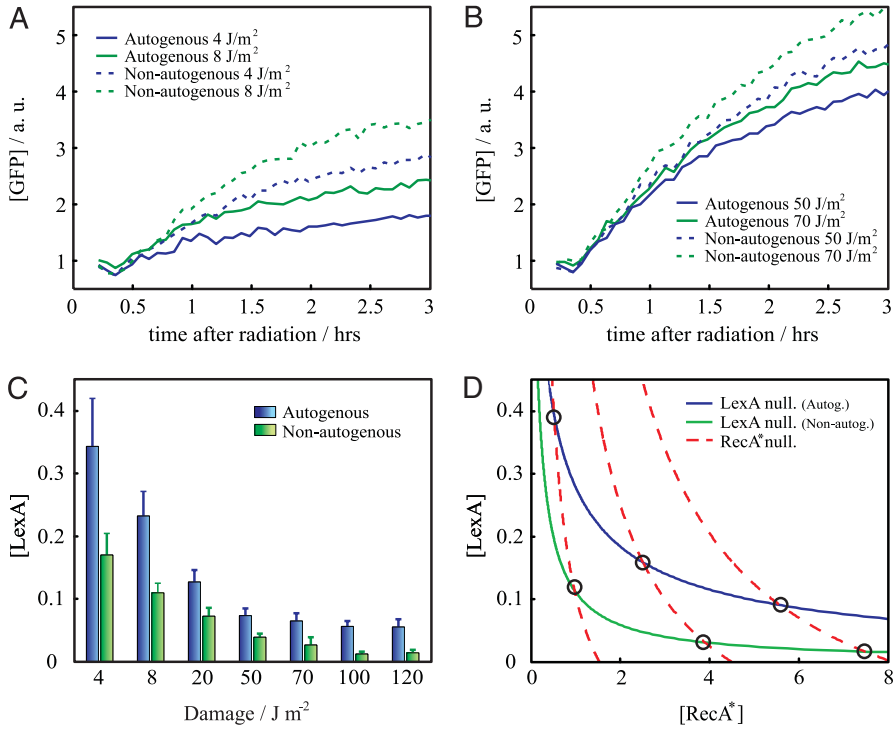


Fig. 3. SOS response to a wide range of inducing signal. (*A* and *B*) SOS dynamics (GFP fluorescence) after UV irradiation in cells with (solid line) or without (dashed line) autogenous regulation in a regime of low (*A*) or high (*B*) damage. Overresponse is stronger at low-UV dose. Reproducibility error of experiments performed on different days is ≈10%. (*C*) Minimal LexA (repressor) concentration as a function of UV dose for the autogenously (blue) and nonautogenously (green) controlled circuits. (*D*) [LexA]–[RecA*] (activated RecA) phase space. Circles denote equilibrium states. The more proportional response of the autogenous system corresponds to a more gradual decrease of the LexA nullcline as activated RecA increases. Overresponse of the nonautogenous system is here reflected in the distribution of equilibrium states close to minimal [LexA] values. Parameters are as follows: $\alpha_{au} = 6$, $k = 0.2$ (see *Materials and Methods*), $k' = k$ (36), $\beta = 1.5$ (36), $\delta_x = 1$, $\delta_y = 1$, $\delta_{xy} = 8$, and $\chi = 0.17, 0.5, 0.9$ (corresponding to increasing RecA* nullclines). Protein concentrations are in units of uninduced LexA equilibrium.

and compare its response with or without the presence of this control on its master regulator, the protein LexA.

We first studied how autogenous regulation determines the strength of the SOS response. Under DNA-damage conditions, the concentration of LexA is reduced. This change in LexA levels releases a battery of SOS genes linked to several repair processes leading to the recovery of the system to its uninduced state (24, 32). We followed the course of this response by measuring the dynamics of a GFP reporter protein under the control of the *lexA* promoter. For a fixed UV dose, the response of the nonautogenous system is larger than that found in the autogenous one, with this difference being more dramatic at low-UV doses. This difference is a direct consequence of the action of the autogenous control, which counterbalances more effectively the degradation of LexA levels induced by the damage. We partially interpreted this result in terms of the stronger stability exhibited by the autogenous system (20, 22). The ratio of stability between both circuits was found to be close to its theoretical maximum, which indicates that this property is being fully exploited in this network. This aspect has an important biological implication, because it acts as a buffering mechanism that prevents unwanted triggering of the response.

The presence of autogenous control has been proposed to provide faster response dynamics in inducible systems regulated by a repressor (20, 21, 23). Here, we were interested in confirming the presence of such speed-up in a natural network. To this aim, we compared the normalized PA of both circuits after exposure to UV radiation. We found that the induction of the response is independent of the type of control and UV dose, because this process is mainly driven by the fast cleavage of the LexA protein, mediated by activated RecA (25). On the contrary, the return from the induced to the uninduced state is circuit dependent. We highlight two different regimes in this case. For low-UV dose, autogenous control does speed up the recovery of the system to the uninduced state, further corroborating the proposed role of this regulation. However, this recovery time equals the nonautogenous case for high-UV doses. We argued that this finding is associated with the longer time required for repair, which, in this regime, determines the recovery time of the system. These two proposed regimes illustrate the relevance of studying simple motifs in their natural context (15). Finally, although we were able to characterize the SOS recovery time by using this type of population-level measurements, it also would be interesting to study how the presence of autogenous control could influence the pattern of activity of the SOS response at the individual cell level (32).

Autogenous control also exhibited a precise fine tuning of the response. We exposed both circuits to a range of UV doses. We showed that the ability of the network to respond proportionally to the level of DNA damage is lost in the nonautogenously controlled system.

All previous discussions contribute to characterize the functional advantages of autogenous regulation in inducible systems controlled by a master repressor. Because induction due to an input signal is associated with repressor elimination, these advantages are specially maximized for low inputs, such as low DNA damage in the case of the SOS system. With sufficiently high inputs, repressor levels are almost emptied with either type of control. This observation suggests that the compensatory mechanisms associated with autogenous regulation lose efficiency, and consequently the difference with respect with nonautogenous control is diminished. When this response comes at a cost for the growth of the bacteria, again

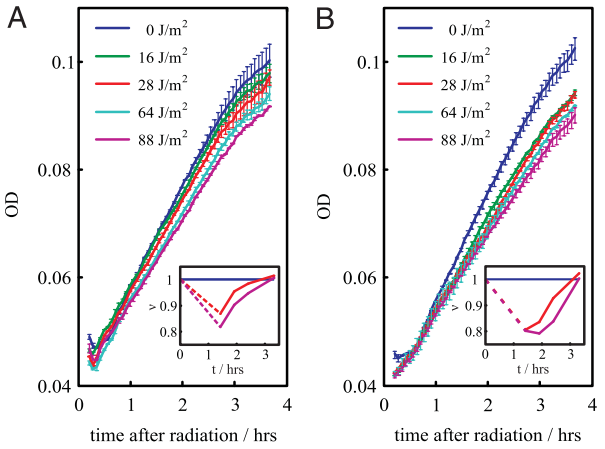


Fig. 4. ODs after irradiation. Exposure to increasing UV dose is reflected in a gradual reduction of growth. This reduction in cells carrying the autogenous circuit (*A*) is smaller than that observed in cells bearing the nonautogenous one (*B*) for an equivalent damage situation. Thus, the overresponse induced by the nonautogenous control comes at a cost. (*Insets*) Growth rate (ν) vs. time for intermediate and high UV doses relative to a nonradiated situation (same color code as main images).

as in the SOS system, the precise regulation associated with the autogenous strategy minimizes such a cost. This is likely to be the case for other networks also exhibiting the single-input module motif where overproduction of the set of operons that are controlled by the single transcription factor would amplify response cost (3). Indeed, in these structures, such master transcription elements are generally autorepressed (3).

These results additionally hint at the reasons why the presence of autogenous control could be selected in some genetic networks. Nature environments are rarely constant. Under these fluctuating/variable pressures, the adaptive response originated by the autogenous control mechanism implies the possibility of achieving high fitness in a range of environments. This adaptive response is favored over specialization, e.g., constant strong response in our case, which would likely cause a higher fitness but only in a single scenario (33). Finally, a plausible negative control of these hypotheses is the nonautogenous regulation of the *lexA* gene recently found in *Leptospira interrogans* (34). These parasitic bacteria are continuously subjected to strong DNA-damaging host-defense factors. In this environment, the disadvantages caused by the lost of autogenous regulation are compensated for by the need of a continuous pool of repair proteins.

In summary, we have characterized the benefits of autogenous control in a natural network. This control structure acts as an optimal strategy to respond to multiple levels of the input signal, minimizing in this way the costs of the response.

Materials and Methods

Bacterial Strains and Plasmids. *E. coli* strains used in this work were JL794 *lexA*+*sulA*11Δ(*lacI*POZYA)169*rpsL*31 (35) and its Δ*lexA*300 derivative, JL2101. Plasmid pSC101-P_{lexA}::GFP (Kan^R) (36) harbors the *gfp* gene under the control of the promoter of *lexA* gene. Plasmid pJWL70 (35) is a pBR322 derivative harboring the *lexA* gene under the control of the P_{lac} promoter. Plasmid pBR322 (37) was used as control plasmid lacking *lexA*. Plasmid pMMB207 (Cm^R) (38) provided the LacI repressor with the P_{lacIq} promoter. Plasmids pSC101-PlexA::GFP, pBR322, and pMMB207 were introduced by transformation into strain JL794, giving strain FMC110 (autogenous control). Plasmids pSC101-P_{lexA}::GFP, pJWL70, and pMMB207 were introduced into strain JL2101, giving strain FMC120 (nonautogenous control).

Growth and Expression Measurements. Parallel cultures of FMC110 and FMC120 (5 ml) were grown overnight in LB medium with kanamycin (50 μg/ml), tetracycline (10 μg/ml), chloramphenicol (60 μg/ml), and IPTG (3×10^{-5} M) at 37°C with shaking. The preceding IPTG concentration was used to equal the LexA levels of both circuits in absence of damage. These cultures were diluted down to $OD_{590} = 0.01$ (as measured in a Victor2 multiwell fluorimeter; PerkinElmer, Wellesley, MA) in M9 medium supplemented with 0.1% (wt/vol) casaminoacids, 0.4% (vol/vol) glycerol, 2 mM $MgSO_4$, 0.1 mM $CaCl_2$, and 0.05% (wt/vol) vitamin B_1, under the set of conditions initially mentioned. To measure growth and expression, we followed a described protocol (36). Briefly, cultures were grown in the fluorimeter at 37°C with 30 sec of shaking at intervals of 4 min. Cultures were irradiated with a UV lamp (Model VL-6C; Vilbert–Lourmat, Torcy, France) ($\lambda = 254$ nm) after reaching $OD_{590} = 0.04$. The plate was then returned to the fluorimeter, after addition of 100 μl of mineral oil per well to prevent evaporation. A second repeated protocol was implemented that included shaking (2 mm orbital, normal speed, 30 sec), absorbance (OD) measurements (590-nm filter, 0.5 sec), and fluorescence readings (filters 485 nm, 535 nm, 0.5 sec, continuous-wave lamp energy setting 10,000). Time between repeated measurements was 4 min. GFP protein concentration was computed by dividing fluorescence by absorbance measurements. We calculated PA (also see *Mathematical Modeling*) by first fitting the GFP/OD data to a sixth-order polynomial and then taking the derivative of such curves (36). To avoid possible artefacts due to the polynomial fitting, we used as an alternative a regression spline method. Both methods showed very similar results.

Mathematical Modeling. We introduced simple mathematical models to study several aspects of the SOS response. These models only incorporate the essential molecular constituents of the SOS network to help compare the behavior of autogenous and nonautogenous control. To understand the difference in stability of the uninduced equilibrium state, we described LexA dynamics in both circuits as $dx/dt = \Pi(x) - \delta_x x$, with $\Pi(x) = \alpha_{au}/(1 + x/k)$, or $\Pi(x) = \alpha_{na}$, for the autogenous (au) and nonautogenous (na) system, respectively (20, 23). Here, α_{na} is the constant PA of the nonautogenous circuit, α_{au} is the PA in absence of repressor of the autogenous one, k is the dissociation constant, δ_x is the protein degradation rate, and x denotes LexA concentration. LexA concentration in equilibrium for the nonautogenous system is given by $x_{eq} = \alpha_{na}/\delta_x$. Thus, to fulfill the condition of same equilibrium state in both circuits, one gets $\alpha_{au} = \alpha_{na}(1 + x_{eq}/k)$. We can visualize a first-order system $dx/dt = F(x)$ as a heavily damped particle inside a potential well. Under this formalism, $F(x) = -dV(x)/dx$. Integrating the dynamics of both circuits, we obtain the potentials of Fig. 1*D*. The stability of the systems is given by the value of the second derivative of this potential with LexA concentration in equilibrium $\sigma = dF(x)/dx|_{x=x_{eq}} = -d^2V(x)/dx^2|_{x=x_{eq}}$ (26).

One can easily obtain the ratio of stabilities for both systems ($\theta = \sigma_{au}/\sigma_{na}$) as

$$\theta = 2 - \frac{\alpha_{na}}{\alpha_{au}}. \qquad [1]$$

This ratio is limited to values ranging from $\theta = 1$ (lack of autogenous control) to $\theta = 2$ (strong autogenous control). Would the presence of delays related to the formation of active repressor modify these stability arguments? This hypothesis could be the case when protein production during delay time (T) was of the order of LexA concentration in steady state (23). For large LexA fluctuations of $\approx$10% of its equilibrium value, significant effects on stability would be observed for delays of the order of the inverse of the LexA degradation rate, i.e., $T \approx 1/\delta_x = \tau/\ln 2 > 1$ h, with LexA half-life

$\tau \approx 60$ min (25). However, delays of the order of only a few minutes are expected in bacterial systems (23).

To analyze how the strength of the response differs for each circuit, we introduced a two-dimensional model describing the dynamics of LexA and activated RecA. The transition from RecA to RecA* is assumed to be in equilibrium, and in this way, RecA* activity becomes proportional to RecA dynamics

$$\frac{dx}{dt} = \Pi(x) - \delta_x x - \delta_{xy} xy, \qquad [2]$$

$$\frac{dy}{dt} = \frac{\chi\beta}{1 + x/k'} - \delta_y y. \qquad [3]$$

Here x and y denote LexA and activated RecA concentrations, respectively; $\Pi(x)$ is the LexA circuit-dependent PA as before; β is the unactivated RecA PA in the absence of repressor; k' is the dissociation constant of LexA repressor with respect to the RecA promoter region; and δ_x and δ_y are the degradation rates of LexA and activated RecA, respectively. In addition, χ is the ratio of RecA activation that is proportional to DNA damage (ssDNA). This value ranges from 0 (no damage) to 1 (virtually all RecA turns into activated RecA). LexA cleavage is included as a term proportional to LexA and activated RecA concentrations with rate δ_{xy} (25).

By considering a situation with a fixed χ parameter, we can study the initial course of the response and, in particular, the distribution of induced LexA equilibrium states as a function of the amount of damage. This description is of course an approximation of the more complicated dynamics (39) of the response but corresponds to a realistic scenario in which LexA cleavage dominates such initial dynamics (25). The difference in response for both systems can be analyzed by using the associated response curves (nullclines) in the phase plane (26).

Both systems share the y-equation (Eq. **3**), and thus the corresponding activated RecA nullcline ($dy/dt = 0$),

$$y(x) = \frac{1}{\delta_y} \frac{\beta\chi}{1 + x/k'}. \qquad [4]$$

This equation constitutes a χ-parametric family of curves whose location is displaced toward low activated RecA as damage decreases, collapsing toward the x axis in the absence of damage, $\beta = 0$ (note that axes were swapped in Fig. 3 *C* and *D*). On the other hand, the LexA nullcline ($dx/dt = 0$) is different for each system:

$$y(x)_{au} = \frac{1}{\delta_{xy}}\left[\frac{\alpha_{au}}{x(1 + x/k)} - \delta_x\right],$$
$$y(x)_{na} = \frac{1}{\delta_{xy}}\left(\frac{\alpha_{na}}{x} - \delta_x\right). \qquad [5]$$

Here, the rate between the first terms of the previous equations increases from 1 (same uninduced states) to α_{au}/α_{na} when $x \to 0$.

Determination of Parameters. We specified the effective kinetic parameters of both circuits by computing the corresponding PA, which is proportional to the number of GFP proteins present in the system (36): $d[GFP]/dt = \alpha_{au}/(1 + [LexA]/k) - \delta_{[GFP]}[GFP] \equiv \Pi([LexA]) - \delta_{[GFP]}[GFP]$. Here the degradation term is mainly due to dilution by cell growth because the GFP used is very stable. We parameterized α_{au} (the PA in absence of repressor of the autogenous circuit) by using maximal PA values measured under high-UV conditions. We used the data obtained in cells carrying the non-autogenous system, because LexA levels are more effectively emptied in this case. From these values and the relation between both parameters in equilibrium, we find α_{na}. Finally, by normalizing LexA concentrations by their uninduced value, we derived the magnitude of k in the same units. Specific values used throughout the work are as follows: $\alpha_{au} = (6.0 \pm 0.7)\alpha_{na}$, $k = 0.20 \pm 0.03$. Finally, we computed relative repressor concentration dynamics from PA activity as $[LexA]/[LexA]_{eq} = k(\alpha_{au}/\Pi(t) - 1)$. We thus obtained minimal LexA concentrations (Fig. 3*C*) from maximal PA values for the corresponding strain and damage conditions.

We thank José María Gómez-Gómez, Oscar Rueda, and Raúl Guantes for very useful conversations; Uri Alon (Weizmann Institute, Rehovot, Israel), Víctor de Lorenzo (Spanish National Biotechnology Centre, Madrid, Spain), and John Little (University of Arizona, Tucson, AZ) for plasmids and strains; and especially Fernando Blanco for technical support. This work was supported in part by a Ministerio de Educación y Ciencia (MEC) Formación de Profesorado Universitario fellowship (F.M.C.), MEC Grant BFU2004-00879 (to J.B.), and the Ramón y Cajal Program (J.F.P.).

1. Bhalla, U. S., Ram, P. T. & Iyengar, R. (2002) *Science* **297,** 1018–1023.
2. Xiong, W. & Ferrell, J. E., Jr. (2003) *Nature* **426,** 460–465.
3. Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002) *Nat. Genet.* **31,** 64–68.
4. Lee, T. I., Rinaldi, N. J., Robert, F., Odom, D. T., Bar-Joseph, Z., Gerber, G. K., Hannett, N. M., Harbison, C. T., Thompson, C. M., Simon, I., *et al.* (2002) *Science* **298,** 799–804.
5. Barkai, N. & Leibler, S. (1997) *Nature* **387,** 913–917.
6. Little, J. W., Shepley, D. P. & Wert, D. W. (1999) *EMBO J.* **18,** 4299–4307.
7. Barkai, N. & Leibler, S. (2000) *Nature* **403,** 267–268.
8. Kalir, S., McClure, J., Pabbaraju, K., Southward, C., Ronen, M., Leibler, S., Surette, M. G. & Alon, U. (2001) *Science* **292,** 2080–2083.
9. Mangan, S., Zaslaver, A. & Alon, U. (2003) *J. Mol. Biol.* **334,** 197–204.
10. Mangan, S. & Alon, U. (2003) *Proc. Natl. Acad. Sci. USA* **100,** 11980–11985.
11. Kalir, S. & Alon, U. (2004) *Cell* **117,** 713–720.
12. Hooshangi, S., Thiberge, S. & Weiss, R. (2005) *Proc. Natl. Acad. Sci. USA* **102,** 3581–3586.
13. Kollmann, M., Lovdok, L., Bartholome, K., Timmer, J. & Sourjik, V. (2005) *Nature* **438,** 504–507.
14. Guantes, R. & Poyatos, J. F. (2006) *PLoS Comp. Biol.* **2,** e30.
15. Eichenberger, P., Fujita, M., Jensen, S. T., Conlon, E. M., Rudner, D. Z., Wang, S. T., Ferguson, C., Haga, K., Sato, T., Liu, J. S. & Losick, R. (2004) *PLoS Biol.* **2,** e328.
16. Brandman, O., Ferrell, J. E., Jr., Li, R. & Meyer, T. (2005) *Science* **310,** 496–498.
17. Zhou, K. & Doyle, J. (1998) *Essentials of Robust Control* (Prentice–Hall, Englewood Cliffs, NJ).
18. McAdams, H. H., Srinivasan, B. & Arkin, A. P. (2004) *Nat. Rev. Genet.* **5,** 169–178.
19. Dekel, E. & Alon, U. (2005) *Nature* **436,** 588–592.
20. Savageau, M. A. (1974) *Nature* **252,** 546–549.
21. Wall, M. E., Hlavacek, W. S. & Savageau, M. A. (2004) *Nat. Rev. Genet.* **5,** 34–42.
22. Becskei, A. & Serrano, L. (2000) *Nature* **405,** 590–593.
23. Rosenfeld, N., Elowitz, M. B. & Alon, U. (2002) *J. Mol. Biol.* **323,** 785–793.
24. Friedberg, E. C., Walker, G. C., Siede, W., Wood, R. D., Schultz, R. A. & Ellenberger, T. (2006) *DNA Repair and Mutagenesis* (ASM, Washington, DC).
25. Sassanfar, M. & Roberts, J. W. (1990) *J. Mol. Biol.* **212,** 79–96.
26. Strogatz, S. H. (2000) *Nonlinear Dynamics and Chaos: With Applications in Physics, Biology, Chemistry and Engineering* (Perseus, Cambridge, MA).
27. Acar, M., Becskei, A. & van Oudenaarden, A. (2005) *Nature* **435,** 228–232.
28. Brent, R. (1982) *Biochimie* **64,** 565–569.
29. Opperman, T., Murli, S., Smith, B. T. & Walker, G. C. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9218–9223.
30. Csete, M. E. & Doyle, J. C. (2002) *Science* **295,** 1664–1669.
31. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. (1999) *Nature* **402,** C47–C52.
32. Friedman, N., Vardi, S., Ronen, M., Alon, U. & Stavans, J. (2005) *PLoS Biol.* **3,** e238.
33. Dobzhansky, T. (1950) *Sci. Am.* **38,** 209–221.
34. Cune, J., Cullen, P., Mazon, G., Campoy, S., Adler, B. & Barbe, J. (2005) *J. Bacteriol.* **187,** 5841–5845.
35. Little, J. W. & Hill, S. A. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 2301–2305.
36. Ronen, M., Rosenberg, R., Shraiman, B. I. & Alon, U. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 10555–10560.
37. Bolivar, F., Rodriguez, R. L., Greene, P. J., Betlach, M. C., Heyneker, H. L. & Boyer, H. W. (1977) *Gene* **2,** 95–113.
38. de Lorenzo, V., Eltis, L., Kessler, B. & Timmis, K. N. (1993) *Gene* **123,** 17–24.
39. Aksenov, S. V., Krasavin, E. A. & Litvin, A. A. (1997) *J. Theor. Biol.* **186,** 251–260.

Apéndice C

What determines the assembly of transcriptional network motifs in *Escherichia coli*?
(*Artículo* **y** *Suplemento*)

OPEN ACCESS Freely available online

PLoS one

What Determines the Assembly of Transcriptional Network Motifs in *Escherichia coli*?

Francisco M. Camas, Juan F. Poyatos*

Logic of Genomic Systems Laboratory, Spanish National Biotechnology Centre, Consejo Superior de Investigaciones Científicas (CSIC), Madrid, Spain

Abstract

Transcriptional networks are constituted by a collection of building blocks known as network motifs. Why do motifs appear? An adaptive model of motif emergence was recently questioned in favor of neutralist scenarios. Here, we provide a new picture of motif assembly in *Escherichia coli* which partially clarifies these contrasting explanations. This is based on characterizing the linkage between motifs and sensing or response specificity of their constituent transcriptional factors (TFs). We find that sensing specificity influences the distribution of autoregulation, while the tendency of a TF to establish feed-forward loops (FFLs) depends on response specificity, i.e., regulon size. Analysis of the latter pattern reveals that coregulation between large regulon-size TFs is common under a network neutral model, leading to the assembly of a great number of FFLs and bifans. In addition, neutral exclusive regulation also leads to a collection of single input modules -the fourth basic motif. On the whole, and even under the conservative neutralist scenario considered, a substantial group of regulatory structures revealed adaptive. These structures visibly function as fully-fledged working units.

Citation: Camas FM, Poyatos JF (2008) What Determines the Assembly of Transcriptional Network Motifs in *Escherichia coli*?. PLoS ONE 3(11): e3657. doi:10.1371/journal.pone.0003657

Editor: Mark Isalan, Center for Genomic Regulation, Spain

Received September 19, 2008; **Accepted** October 20, 2008; **Published** November 6, 2008

Funding: This work was supported in part by a Ministerio de Educacion y Ciencia (MEC) Formacion de Profesorado Universitario fellowship (F.M.C.), and MEC Grant FIS2006-10368 (to J.F.P.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: jpoyatos@cnb.csic.es

Introduction

The collection of transcriptional interactions in a cell constitutes a network able to sense diverse biochemical signals and execute, in response, a range of cellular programs. Recent analyses of this network revealed a series of strategies of cellular control at the system-level, which have later shown to be applicable to other classes of biological networks [1].

More specifically, the successive analysis of *Escherichia coli*'s transcriptional network, where interactions involve a pair of operons encoding the transcription factor (TF) and regulated genes [text S1 section 1], respectively [2], identified the presence of a number of recurrent regulatory patterns as basic constituents of the network. Initial studies first found the presence of the simplest of these patterns, the one-element feedback loop [3]. More exhaustive examinations confirmed the prevalence of these structures [approximately 56% of *E.coli*'s TFs are autoregulated, Materials and methods], and further observed the use of other types of regulatory circuits, generally termed as network motifs [4].

What do motifs emerge? Two general models are currently considered. The most accepted one associates the presence of motifs to the singular information-processing tasks they can accomplish (see [5], for a review). Additional properties could further support this picture, such as the strong dynamical stability exhibited by motifs [6], or the correlation of their abundance with the global functional requirements acting on the network, e.g., the necessity of short response times in transcription [7]. Motifs in this model are then adaptive and isolated working units, a view that seems partially confirmed by their appearance in several transcriptional networks (e.g., those of *Bacillus subtilis* [8] or *Saccharomyces cerevisiae* [9]), and by the experimental confirmation of some of their suggested functional attributes [10,11,12,13,14].

An alternative model proposes that the occurrence of motifs is rather nonadaptive. Motifs might arise, according to this hypothesis, by the action of neutral population-based forces –like random genetic drift [15]– as the result of intrinsic mechanisms of genome evolution [16,17,18], or as a consequence of null network growth constraints [19]. These aspects would additionally suggest a fuzzy signal of motif conservation across species, a prediction that seems partially confirmed [20]. Moreover, this interpretation also challenges the relevance of motifs as separated functional entities [21].

Here we propose an integrative approach to understand the assembly of motifs that partially solves this controversy. This strategy is based on characterizing the relation between motif assembly and the capacity of their constituent TFs to integrate and transmit biochemical signals. We thus denote the capacity to integrate several environmental stimuli as sensing specificity (coarsely quantified with the presence/absence of upstream transcriptional regulation on a TF which could also be described as sensing ability), and the specificity of the TF to transmit signals as response specificity (this being quantified by the size of the corresponding regulon). We show how these measures –even as coarse as they are– are helpful to obtain a new picture of motif assembly.

In particular, while analysis of the first feature reveals an uneven distribution of autoregulation, the study of response specificity uncovers a linkage between the tendency of a TF to establish FFLs and its regulon size, with the first decreasing with the second. Investigating this pattern in detail, we identify several causes of motif emergence.

First, TFs with small regulons correspond to a class of FFLs caused by the hierarchical regulation of groups of operons, mostly associated to catabolite repression. In comparison, TFs with large regulons (hubs) leads to the assembly of both FFLs and bifans aggregates by coregulating third elements in combination with other hubs. Interestingly, most of this coregulatory signal appears to be neutral following a network null model, which in turn helps us to strengthen the adaptive nature of a complementary small group of such aggregates. Hubs also exhibit a complementary regulatory strategy, i.e., exclusive regulation. This induces the emergence of large single input modules (SIMs) structures, whose appearance is again partially neutral. The basic idea of network motifs was that a number of regulatory patterns appeared in extant networks much more often than in randomized ones [4,22]. Our analysis ultimately shows how only a small subset of motifs, within each motif class, originates the statistical signature that helped unravel these structures in *E. coli*'s transcriptional network.

Results and Discussion

Specificity and autoregulation

In analyzing why autoregulation, the simplest motif, is such a pervasive regulatory attribute in *E. coli*'s network, we could be asking two complementary questions. We could first ask whether autoregulation is usually acting in combination with other transcriptional interactions. This strategy could enhance the interpretation of environmental states [22] by allowing the integration of several signals, i.e., the bacterial sensing specificity [23]. A second question would be whether the distribution of autoregulation relates to the specificity of the response. We followed here a simple network-based definition, and roughly quantified this specificity by the number of genes regulated by the TF (regulon size), e.g., small regulons indicating highly specific responses.

To study the first question, we partitioned all TFs in the network into two broad classes: TFs that do not experience any upstream regulation and those which do. Note that TFs of the first group are at the top of the network multi-layered structure [24,25,26,27] – autoregulation, when present, would act in isolation– while those in the second class constitute the network lower layers. In this latter group autoregulation would act in combination with those TFs exerting upstream regulation (Fig. 1.A, top). We observed a smaller incidence of autoregulated TFs (ATFs) at the top (27 of 63 TFs are ATFs, 43%) as compared to lower layers (37/60, i.e., a 62% with $p = 0.03$ by assigning randomly all autoregulations, i.e., keeping fixed the number of ATFs and network hierarchy, 10 000 times, text S1 section 2).

To examine the second question, we introduced three TF classes in terms of response specificity (quantified by regulon size, Fig. 1.A, bottom). We observed that the tendency to be autoregulated grows with regulon size (Fig. 1.B). Moreover, both sense and response specificity could underlie selection for autoregulation. For instance, a large regulon size TF (a hub) at the top of the network could sense very general nutrient conditions and react by globally changing bacterial physiology [26,27]. One could hypothesize that autoregulation in this case would contribute to a more precise control of the expression of the TF inducing such major physiological changes [22] (see also following discussions).

We thus studied sense and response specificity in combination. We found that the set of TFs lacking upstream regulation and with low regulon size are hardly autoregulated (Fig. 1.C). However, within this same set, hubs are mostly autoregulated (7 out of 8 hubs are ATF, e.g., CRP). These patterns are not observed in TFs under upstream control. In this case, operons exhibited a relatively homogeneous presence of autoregulation, independent of response specificity (Fig. 1.D).

Autoregulation and the assembly of complex motifs

While a relation between autoregulation and regulon size (response specificity) was apparent in some of the previous

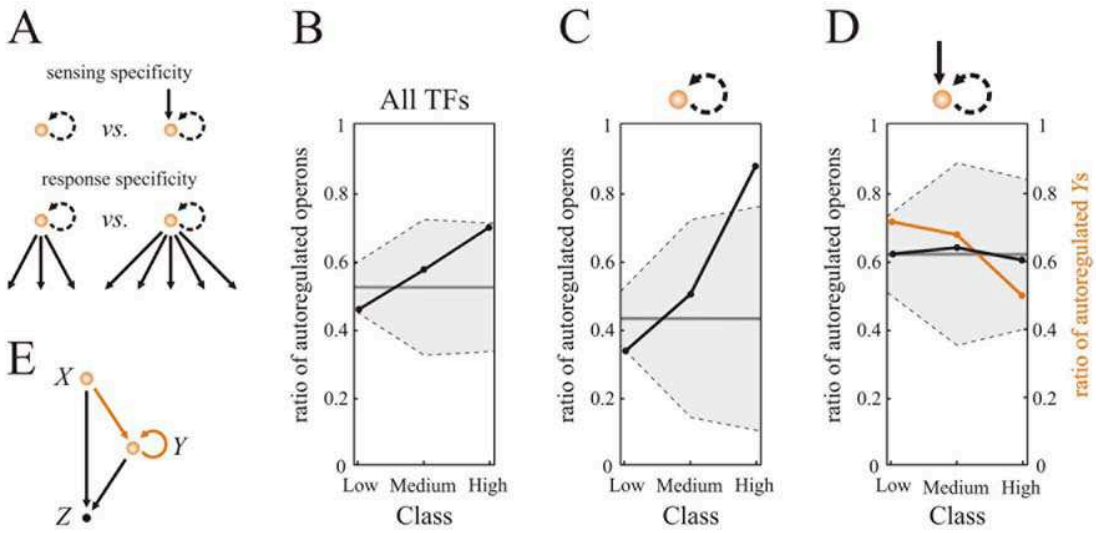


Figure 1. Distribution of autoregulation. (A) Sensing specificity (presence/absence of external regulation) and response specificity (regulon size). (B–D) Abundance of autoregulation in three TF response specificity classes quantified by regulon size. This includes first an analysis including all network TFs (B), followed by two more examinations considering those subsets of TFs without (C) or with (D) upstream transcriptional control, respectively. Classes: low (one to four regulated operons), medium (five to nine) or high (ten or more, considered as TF hubs). We also plotted the null behavior obtained by random sampling of the corresponding class –preserving group size– within the specific TF group (B,C, or D), 10 000 times (mean, continuous gray line, ±2 standard deviations, shaded area). Lines between points to help visualization. (E) Autoregulated TFs with upstream regulation as part of a FFL, a motif constituted by three elements *X*, *Y*, and *Z*, two of them being always a TF in this context (*X* and *Y*). In (D), the ratio of autoregulated *Y* for each specificity class is also showed (orange), see main text.
doi:10.1371/journal.pone.0003657.g001

patterns, the homogeneous distribution of ATFs in lower layers alternatively suggested a fairly neutral linkage between these properties. Could this distribution be masking some other patterns of network organization? Interestingly, ATFs with upstream regulation are common constituents of FFLs [5], with the ATF and the additional regulator as *Y* and *X* of this motif, respectively (Fig. 1.E), so we asked if this association could reveal any pattern.

To investigate this, we initially counted the number of FFLs with an *Y*-element belonging to each of the TF response-specificity classes, and within these groups the percentage of FFLs with autoregulated *Y*s. This revealed a striking dependence with specificity, ranging from a 71% of *Y*s autoregulated in the low regulon-size class to a 50% in the high class (Fig. 1.D, orange line).

To better understand this dependence, we quantified the tendency of a TF (autoregulated or not; but with upstream regulation) to establish FFLs with a measure that we named the FFLness (F, $0 \leq F \leq 1$). This score is the ratio between the number of FFLs with this TF –as *Y*– found in *E.coli*'s network, and the maximum number of FFLs that such TF could potentially assemble (given by the product of the number of upstream TFs regulating *Y* and its regulon size, Fig. 2.A, and Text S1 section 2 and Figs. S1, S2, S3, S4). Figs. 2.B–C shows FFLness as a function of response specificity for autoregulated and non-autoregulated TFs, both in the extant network and in a null model (considering randomized networks with the same connectivity sequence, Materials and methods).

The distribution of autoregulation discussed before is confirmed in the new FFLness measure. For instance, the excess of autoregulation in the low regulon-size *Y*s (Fig. 1.D, orange line, low class) is also reflected in a stronger mean FFLness observed in the low class ATFs with respect to the non-autoregulated ones (compare mean FFLness values of low class in Fig. 2.B and Fig. 2.C, respectively). In addition, the FFLness measure highlighted a characteristic decay with regulon size in the establishment of FFLs by a TF, not found in the null model (Figs. 2.B–D, shaded areas). Since this decrease is observed in both autoregulated and non-autoregulated TFs, and since the distribution of autoregulation is, in contrast, homogeneous with regulon size (Fig. 1.D, black line), we asked what additional factors could help explain the strong tendency to establish FFLs (high FFLness) and its decay.

What underlies the strong tendency to establish FFLs?

In the following, we considered two neutral models that could contribute to the strong tendency of low regulon-size TFs to establish FFLs: genomic architecture and homology between motif constituents. For the first model, we examined the association between this signal and neighbor regulation –of a TF on a genomically adjacent operon– a characteristic genome architecture in prokaryotic transcriptional control [28,29] that can readily promote FFL assembly (Fig. 2.E). However, genome architecture could only partially explain high FFLness, as low regulon-size TFs with upstream regulation also showed a strong disposition to establish FFLs with nonadjacent operons (Figs. S1–S2) that did not even colocalize in the genome in broader terms (text S1 section 4). We thus analyzed if homology between motif components could be explaining this pattern [32,33]. The fact that these combined models could not totally account for the strong FFLness score, and that we also identified a remarkable functional association between the constituents of these FFLs, indicates selection for a pattern of aggregated FFLs that we propose in a later section in detail.

Is genome architecture driving high FFLness? In Fig. 2.D we plotted the proportion of neighbor regulation in TFs under upstream control. Low regulon-size TFs are indeed enriched by this architecture, a signal that decreases with regulon size (similar to FFLness, also in Fig. 2.D). This suggests neighbor regulation as

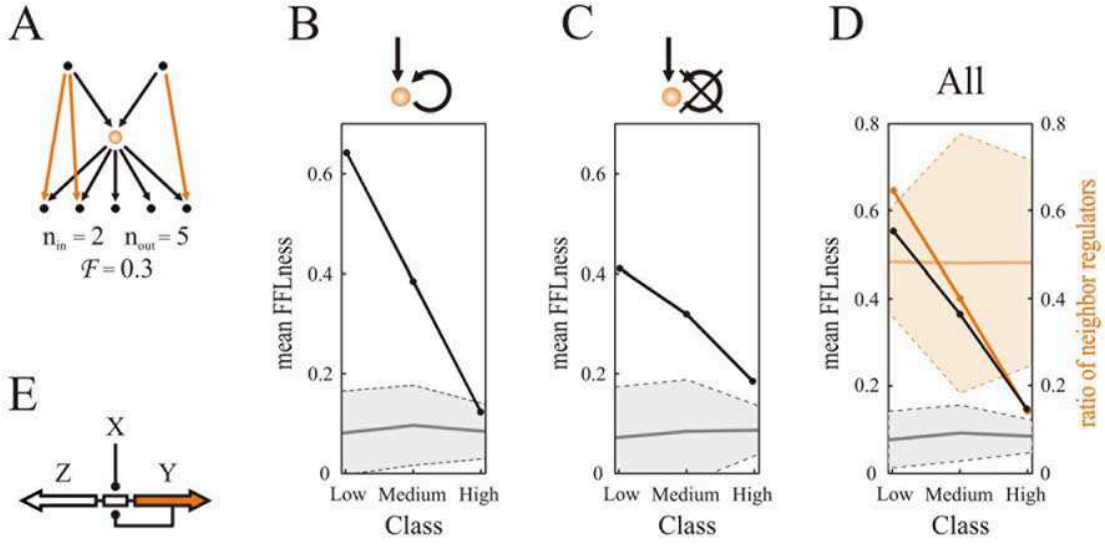


Figure 2. Assembly of FFLs and TF regulon size. (A) Computing FFLness: the maximum number of FFLs that can be potentially assembled by the TF in this example is $n_{in}n_{out} = 10$. Imagine that only 3 FFLs were actually observed (orange arrows), then this TF would have $F = 0.3$. (B–D) Mean FFLness as a function of regulon size for TFs with upstream control (classes defined as in Fig. 1). (B–C) autoregulated/non-autoregulated TFs. (D) all TFs with upstream control. The null behavior obtained in a network null model (Materials and methods; mean, continuous gray line, ± 2 standard deviations, shaded area) is also plotted. In (D) we additionally showed the ratio of TFs –with upstream control– regulating a neighbor operon (dark orange) and its corresponding null (mean, continuous light orange line, ± 2 standard deviations, orange-shaded area). Lines between points to help visualization. (E) Divergent architecture could promote FFL assembly. Spacer between two divergent operons (*Y* and *Z* in FFL) could include a binding site leading to the coregulations of these operons by an upstream TF (*X*) and by the autoregulated TF (*Y*). Note how this genomic architecture links autoregulation to neighbor regulation [30,28,29,31].
doi:10.1371/journal.pone.0003657.g002

an important factor underlying part of the high FFLness signature. Indeed, although autoregulation appeared linked to high FFLness (but not significantly, ATFs: $F = 0.64$; non-autoregulated TFs: $F = 0.41$, $p = 0.12$, Wilcoxon rank sum test; Figs. 2.B–C, low regulon-size class), neighbor regulation is a stronger determinant (TFs with adjacent regulation: $F = 0.70$; TFs without: $F = 0.29$, $p < 0.01$, Wilcoxon rank sum test).

In this same analysis, we also recovered the connection between neighbor control and autoregulation –previously reported [30,28,29,31]– for the TFs at the top of the network ($p = 0.01$, two-tail Fisher's exact test). This relation was lost in those TFs with upstream control ($p = 0.5$, Yates-corrected χ^2-test, see also text S1 section 3 and Table S2). This could be partially caused by a failure to report autoregulation in some cases –which can be particularly difficult to resolve for divergent architectures [31]. Alternatively, the acquisition of new binding sites to enable the upstream control could in some occasions interfere with the autoregulatory binding site.

Is homology driving high FFLness? We considered two possible models additionally contributing to the high FFLness score. The first one explained this tendency by the homology between those TFs encoded in the *X*- and *Y*-operons. Alternatively, a second model analyzed if nonadjacent *Z*s (nad*Z*s) inherited the same regulation of the central unit –which would lead to the establishment of additional FFLs– by duplication of genes belonging to such central set. This unit was defined as the group of genes constituted by the operon encoding the TF acting as *Y* and, when applicable, by those of its *Z*-operons adjacently located (which included also second adjacents, to control for tandem duplications).

We found 15 out of 40 FFLs constituted with nad*Z*s that could be explained with the homology models above (Fig. 3.A. and text S1 section 4). Thus, both null models only partially contributed to explain the strong tendency to assemble FFLs with nad*Z*s, even though we considered very permissive scenarios. For instance, our reasoning in the first model assumed that the duplication of *X* happened after this factor established its regulatory links, a relaxed assumption considering the prevalence of HGT [34] and the high rate of network rewiring in bacteria [35], while the conservation of regulation in the second model did not consider the influence in this conservation of the order of genes on the operons [36].

Is functional fine-tuning driving high FFLness? What about the rest of FFLs that could not be explained by the models above? We observed a characteristic functional pattern based on the following features. First, in most cases CRP is acting as *X* of the FFL (20 out of 25 cases of FFL not explained by homology, Fig. 3.A). While this could be *a priori* expected due to the large regulon of CRP, we found that this role is particularly relevant in these 25 cases within the low regulon-size group ($p < 0.014$, two-tail Fischer's exact test). Note that CRP is also dominant in the latter group as compared with the rest of FFLs ($p = 0.008$, Yates-corrected χ^2-test).

Second, the function of the genes encoded in nad*Z*s is remarkably related to the one exhibited by genes in the central unit. Sometimes genes in nad*Z*s and those encoded in the corresponding central operon are transporters responding to the same metabolite by alternative mechanisms. For example, arabinose and galactose sugars can be imported by low affinity proton-driven MFS symporters or by high affinity ATP-driven ABC transporters, Fig. 4. Other relationships do not necessarily imply different transporter classes. The transcriptional factor DgsA controls three nad*Z*s associated as follows: two of them encode sugar-specific components of respective glucose PTS transporters,

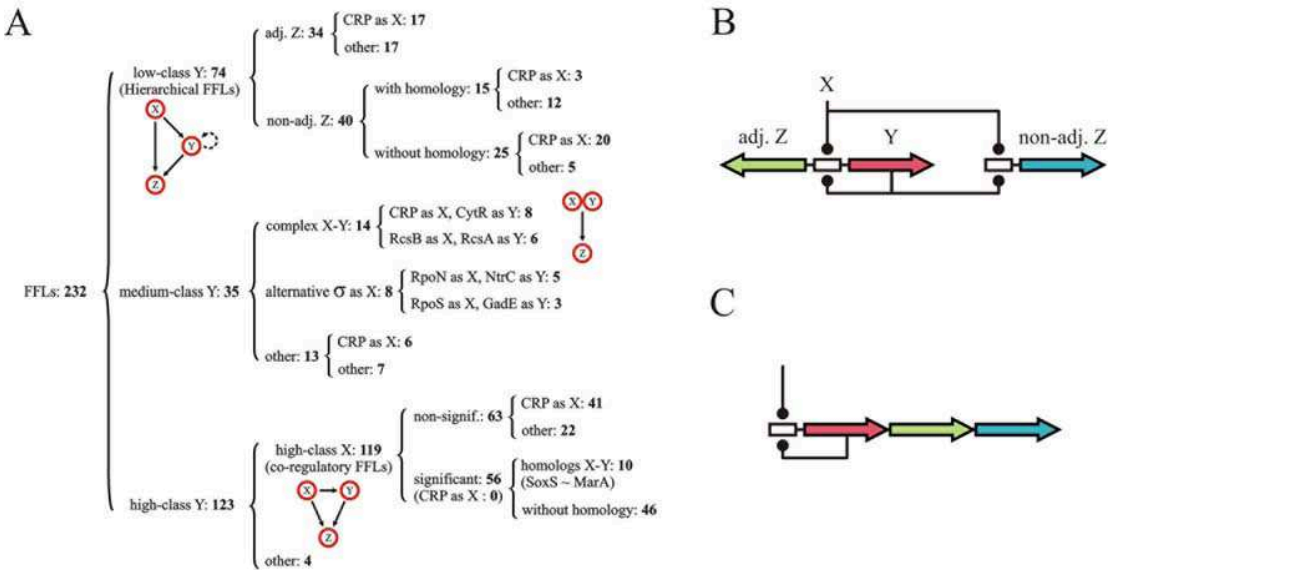


Figure 3. FFL classification. We divided the 232 FFLs identified in the network into three response specificity classes as defined in Fig. 1. i) Low class, hierarchical. These FFLs present *Y*-elements with small regulon size and a tendency for autoregulation. We divided this class in two subfamilies defined by whether *Y* regulates, or not, a genomically adjacent gene: 34 FFLs have an adjacent *Z*, while 40 FFLs do not have an adjacent *Z*. In the first subfamily we scored the number of FFLs with *X* being CRP. In the latter subfamily, we quantified those FFLs exhibiting homology and, again, the role of CRP in these groups. For instance, we found 20 low regulon-size FFLs with CRP as *X*, with *Z* not adjacently located and whose assembly does not follow the homology model. ii) Medium class, complexes. FFLs established with *Y*s of this class are enriched by pairs (*X*,*Y*) in which the action of one TF totally relies on the presence of its partner, e.g., RpoN on NtrC, see Fig. S3). iii) High class, hub coregulation. These FFLs mostly correspond to those motifs whose *Y*-elements are hubs. We again characterized this class with respect to homology and CRP influence. Note that all numbers indicate those FFLs found in each category, see also Table S1. (B–C) Dual regulatory logic in hierarchical FFLs (B) *vs.* polycistronic (C) designs, color code represents functionally equivalent genes, text for details.
doi:10.1371/journal.pone.0003657.g003

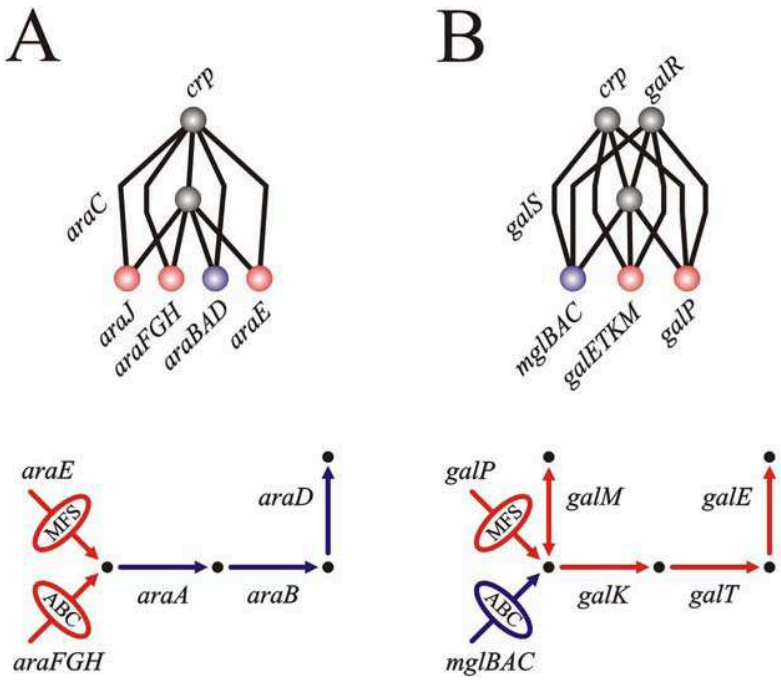


Figure 4. Examples of functional fine-tuning in *Z*s of hierarchical FFLs: the arabinose (A) and galactose (B) systems. In each case we plotted (top) the incoming/outgoing regulatory interactions associated to the TF sensing the specific sugar and those links involved in FFL assembly, and (bottom) the encoded metabolic pathway, with arrows or ellipses –crossed by arrows– denoting enzymes and transporters, respectively. Each system imports the corresponding metabolite by two different (non-homolog) transporter classes: MFS sugar/proton symporters and ABC transporters. MFS transporters encoded in *araE*, *galP* are homologs. ABC transporters in *araFGH*, *mglBAC* are also homologs. Note that both examples exhibited maximal FFLness, i.e., $F = 1$. Color code *Z*-elements: blue, adjacent *Z*; red, non-adjacent *Z*. The same color code applies to the encoded pathway steps. Text S1 for more examples and further discussions.
doi:10.1371/journal.pone.0003657.g004

while a third nad$Z$ encodes the common non-sugar-specific components of this transporter type. Finally, those nad$Z$s encoding, apart from transporters, also enzymatic reactions are all associated to one-step pathways (e.g., *chbBCARFG* in charge of chitobiose degradation) which are complementary to those found in their respective central unit (text S1 section 4 and appendix, Table S10, for details).

Hierarchical aggregated FFLs as adaptive functional units

What is the overall picture suggested by the discussions above? We emphasize here the hierarchical regulatory scheme that we identified, and propose an adaptive scenario for its emergence. This scheme combines the action of a general and specific TF following a hierarchical logic mostly linked to catabolite repression: when glucose is absent, CRP regulation (X in most of these FFLs, Fig. 3.A) activates a number of genes enabling the sensing (Ys in FFLs, usually autoregulated) and metabolizing (Zs, adjacently or not adjacently located) of alternative sugar sources. Y in these FFLs is thus subordinated to X activity, and the control of each group of operons (Zs) by the corresponding (X,Y) hierarchical regulatory logic presents this type of aggregated FFLs as fully-fledged independent functional units. Autoregulation in TFs encoding Y also implies that this logic applies to this very same TF, suggesting that its presence is not just an optional regulatory design, but rather a fundamental ingredient of transcriptional control. The implementation of this hierarchical control might not be necessarily restricted to FFLs (see below).

With respect to the adaptive/neutral forces leading to the assembly of these aggregates, we can envisage the following scenario. Initially, pairs of neighborly regulated genes –in divergent orientation– can be horizontally transferred to *E. coli*, and then acquire an additional regulation by a global regulator, mostly CRP [37], in their intergenic region. This leads indirectly to the assembly of a core FFL which could only be a neutral byproduct of the previous process (Fig. 2.E). This neutral picture of FFL assembly seems to be lost when we consider the other FFLs in the aggregate.

The functional characterization of nad$Z$s revealed a very close relation among all Zs, adjacent or not (see Fig. 4, text S1 section 4 and appendix), and indicates that the emergence of these aggregated FFLs could be a consequence of selection for the (X,Y) combinatorial logic, independent of genome location. This model could be further supported if some similarity of expression and/or evolutionary dynamics between adjacent and non-adjacent Zs were observed. The first one was experimentally reported in the arabinose system (adjacent –*araBAD*– and nonadjacent –*araFGH*– Zs [10]). In addition, we found that the averaged phylogenetic co-conservation of the pairs (Y,Z) in γ-proteobacteria was larger than expected by chance, and that the difference in this co-conservation for adjacent and nonadjacent Zs, studied independently, was non-significant (text S1 section 4).

Moreover, we propose in this context the similarity between hierarchical FFLs and autoregulated polycistrons. Genes acting as Zs in the former would be part of the polycistron in the latter and could also be subjected to a hierarchical logic (by an external global signal and the specific one associated to the polycistronic autoregulation, Fig. 3.B–C). This equivalence is based on the following observations. It is implied by the fact that the set of low regulon-size autoregulated operons, which do not regulate adjacent ones, is enriched with long polycistrons only when exhibiting external regulation (Tables S3, S4, S5). It is also

suggested experimentally. One of the Z-elements of the *gal* system (Fig. 4.B) was showed to exhibit the very same response speed-up to that observed in a negatively autoregulated polycistron [13]. Indeed, the Y-element of this FFL –GalS– is negatively autoregulated, sharing thus the very same transcriptional logic of this Z-element. When to present a hierarchical FFL or an autoregulated polycistron regulatory architecture could be related to the specific mechanisms of network evolution (see text S1 section 4) [38,37].

What underlies the weak tendency to establish FFLs?

Resuming the analysis of the tendency of TFs to establish FFLs, beyond hierarchical FFLs (Fig. 2.D), we first observed that FFLs established with medium regulon-size Ys are enriched by pairs (X,Y) in which the action of one TF totally relies on the presence of its partner, e.g., RpoN on NtrC (see Fig. 3.A and Fig. S3), in order to induce the expression of third operons (Zs) and of Y itself (which is always autoregulated in these cases). This directly leads to the emergence of a FFL structure. What about the drastic decay of FFLness found in TFs with large regulon size?

The decrease in FFLness in these TFs (autoregulated or not) made this signal closer to the one seen in the null model (although remaining statistically significant, Fig. 2.D). Such low value does not imply that these TFs do not establish FFLs –in fact more than half of the FFLs of the network has a hub as Y-element, Fig. 3.A– rather that a small fraction of their potential FFLs are assembled. Is then this family of FFLs mainly originated by neutral forces?

The likelihood of neutral assembly of a FFL by a given TF acting as Y ($\mathrm{FFL}_{\mathrm{null}}$) is given by the product of the neutral FFLness (according to the network null model, i.e., $F_{\mathrm{null}}\approx.08$, Fig. 2.D), the number of external regulations (n_{in}) and the regulon size (n_{out}): $\mathrm{FFL}_{\mathrm{null}} = F_{\mathrm{null}} n_{\mathrm{in}} n_{\mathrm{out}}$. This directly indicates that the null assembly ($\mathrm{FFL}_{\mathrm{null}}$) scales with regulon size, and also that the low FFLness of Y hubs can still be associated to the appearance of a considerable number of FFLs. This is caused by the partial random overlap of the regulons of the potential X and Y TFs, an overlap favored when both TFs are hubs. Indeed, X elements are mostly hubs in the extant network (211 cases of the total of 232; in particular 119 FFLs have both X and Y hubs, Fig. 3.A). Is then neutral overlap a major contributing part of the FFLness score observed among hubs?

Hub combinatorial regulation, FFLs and bifans motifs. We investigated the relevance of neutral hub coregulation as follows. We identified all possible pairs of hubs in *E.coli*'s network (23 hubs, 253 pairs). For each pair, we contrasted the coregulation observed in *E. coli* with a null averaged value obtained with randomized networks (Materials and methods and text S1 section 2). Interestingly, we found a fairly small number of significant coregulations after correcting for multiple testing (Table 1 and text S1 section 2). Note that the most significant ones correspond to five pairs in which one TF regulates the other, i.e., they are associated to FFL aggregates, while the remaining ones –when hubs do not interact– correspond to bifan aggregates [4,39,40] (Fig. 5).

What type of adaptive coregulations revealed this simple null? Gene duplication of regulatory hubs was suggested to play an important role in the assembly of FFLs in yeast [17], so we first analyzed if duplication could be contributing to these significant coregulations. We found that only the (MarA,SoxS) pair showed homology. This common regulation indeed arose by duplication [37]. Notably, in each of the rest of significant interacting pairs a coordinated way of action is particularly well documented. This is the case of the second top (FlhDC,FliA) pair, which is part of the genetic network controlling the temporal program of flagellar assembly, with FlhDC being its principal regulator, and FliA the flagellum-specific σ factor [12]. Additionally, the pairs (FNR,NarL) and (FNR,ArcA) regulate anaerobic respiration and fermentation. In this context, ArcA and NarL determine the type of respiration mode under the coordination of FNR [41]. Finally, RpoE is involved in heat shock and other stress responses [42]. This TF shows a strong tendency to exclusive regulation (40 out of its 51 regulatory outputs do not receive any other transcriptional regulation, see next section). However, reaction to membrane stress is coordinated by coregulation with CpxR, a constituent of the two-component regulatory system CpxA/CpxR, which senses this type of stresses (including misfolded proteins and degrading factors). Moreover, the set of non-significant coregulations ($p>0.1$, even before controlling for multiple testing) and those established by the (MarA,SoxS) pair implied a total of 73 out of 119 FFLs,

Table 1. Pairs of hubs exhibiting significant coregulation.

TF pair		cre	cre_n	Z-score*	FFL/Bi-fan**
SoxS	MarA	10	66.7	17.68	→
FliA	FlhDC	5	40.8	10.52	←
FNR	NarL	18	42.4	9.97	→
FNR	ArcA	16	27.3	5.21	→
RpoE	CpxR	7	21.4	4.74	←
IHF	RpoN	10	22.6	4.58	n.i.
FNR	IHF	18	25.9	4.39	n.i.
IHF	NarL	8	23.3	4.28	n.i.
IHF	Lrp	7	19.9	3.63	n.i.
CRP	ArcA	19	38.2	3.56	n.i.
CRP	RpoE	2	3.6	−3.21	n.i.

TFs in pairs sorted by regulon size. * Cutoff corresponds to adjusted $p<0.05$. ** Arrows denote regulatory order. cre: number of coregulations, n.i.: non-interacting hub pairs. cre_n: normalized cre as $\mathrm{cre}/\sqrt{R_1 R_2}$, R_i denotes regulon size. Note that the pair (CRP, RpoE) appears as the single case of significant anticoregulation (see main text).
doi:10.1371/journal.pone.0003657.t001

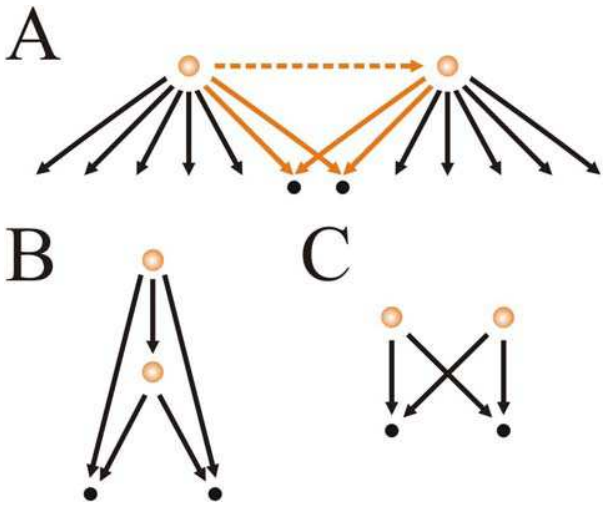


Figure 5. Motif assembly by random overlap of large regulons. (A) In this example a couple of hubs shares two target (coregulated) operons. As hubs may interact or not (dashed line), these coregulations lead to the assembly of aggregated FFLs (B) or bi-fan motifs (C).
doi:10.1371/journal.pone.0003657.g005

with X and Y being interacting hubs, and whose presence could be explained by neutral processes alone according to the null network model considered.

It is also interesting to observe that the mean FFLness of the *Y*-elements being part of the significant FFLs (beyond MarA, i.e., FliA, NarL, ArcA and RpoE) is $F = 0.34$, while a much smaller averaged value was observed for the rest of hubs with upstream regulation and non-significant co-regulations, i.e., $F = 0.07$. This thus helped explain why we found a small, but significant, FFLness in the large regulon-size class: this signal is a mixture of the non-significant and significant coregulations represented in the decayed, but yet significant, total score (Fig. 2.D).

What about those significant coregulations established by non-interacting hubs in Table 1? The integration host factor (IHF) regulator appeared recurrently in this case (4 out 5 cases). This trend could be explained by the intrinsic architectural role of IHF that facilitates the action of other TFs by controlling DNA bending [42]. Interestingly, although there are 11 pairs of homolog hub pairs without mutual regulation, none of them leaded to a significant coregulation. For example, CRP and FNR –the hubs with the largest regulons– are homologs [27], and coregulated a number of operons similar to the null value (observed: 24 operons, random: ~20, $\tilde{Z}$-score = 1.15). The pair (CRP, RpoE) did arise as a single case of significant anti-coregulation, i.e., they coregulated less operons than expected by chance. The autonomy of the RpoE stress response is thus reflected in a necessary uncoupling of the metabolic context.

Since duplication of hubs did not play a relevant role in these significant coregulations [17], we asked if duplications of coregulated genes could contribute to this signal. This could be partially the case in the coregulations established by IHF, FNR and NarL (these hubs shared 8 coregulated operons), and only under a permissive criterion (inheritance of binding sites imposes more strict constrains to the location of homolog genes in their respective operons [36]). However, there are functional arguments for the convergent establishment of these coregulations, since IHF enhances the action of the activators NarL and FNR, as stated above [43].

Exclusive regulation and single input modules

A somehow complementary regulatory strategy to combinatorial regulation is linked to exclusive regulation, this term referred to the absence of any additional regulation on a group of operons, beyond that of a given master TF. We first investigated if exclusive regulation is significantly observed in the extant network (comparing with a null network model, as previously). Note that this question is equivalent to ask whether single input modules (SIMs) are network motifs [4]. We found 27 exclusive regulations (≥3 target operons) or SIMs in *E.coli*'s network. This number is not significantly different to the random score (30.5 SIMs, $p = 0.11$). However, the mean number of target operons per SIM is indeed larger than expected (observed: 10.3 targets, expected: 8.4, $p = 0.0032$), confirming that large SIM structures are network motifs. Could this exclusive regulation be a statistically significant pattern uniquely associated to a small number of SIMs as we found in the case of the hub coregulation signal?

To analyze SIM motifs individually, we computed for each TFs in the network regulating≥3 operons the ratio between the number of operons controlled exclusively and its regulon size –all interactions with same sign, dual regulations not considered. We named this score the SIMness (S, $0 \leq S \leq 1$) of the TF, and compared it to the averaged value obtained in a null network model (text S1 section 2). We discovered only a limited number of large regulon-size TFs with significantly high SIMness (Table 2). These structures are undeniably among the most isolated functional units in the transcriptional regulatory network.

Table 2. Positive and negative SIMness.

TF	R^+	S^+	S_r^+	*Z*-score*
RpoE	51	0.78	0.33	7.34
CRP	117	0.15	0.38	−5.98
Fis	41	0.68	0.33	5.10
RpoH	25	0.68	0.31	4.07
IHF	28	0.07	0.33	−3.01
TF	R^-	S^-	S_r^-	*Z*-score*
LexA	19	0.89	0.31	5.65
Fur	28	0.57	0.31	3.20

*Cutoff corresponds to adjusted $p<0.05$. R^i: regulon size, S^i: SIMness score, S_r^i: random SIMness score, *i*: positive or negative.

doi:10.1371/journal.pone.0003657.t002

What functions are associated to these SIMs? They generally corresponded to autonomous systems able to rapidly induce urgent cellular responses. The SIM for which Fis is the master (positive) regulator is constituted by 28 operons involving a total of 75 genes. These genes are mostly constituted by transfer or ribosomal RNA genes (70 out of 75) coordinately expressed as adaptation to rapid growth conditions [44]. Three of the remaining cases are stress response regulators: LexA, exhibiting the highest (negative) SIMness, controlling DNA damage response [14], RpoE and RpoH, regulating several stresses like those related to heat shock [42] (RpoE also showed a strong coregulation when acting with CpxR, as discussed before). Finally, Fur is in charge of the control of iron homeostasis [45]. Homology is again not relevant in these significant SIMs, like in the case of significant hub coregulations. We only found one case of homology between the master regulator and its targets (LexA,U-muD). Homology among target genes was also rare (data not shown).

Note that Table 2 also included two TFs which displayed significant anti-SIMness, i.e., they regulated exclusively less operons than expected by chance. Anti-SIMness of CRP and IHF are a consequence of their strong bias to coregulation. We argued above that IHF is involved in the assembly of bifans in combination with several hubs. Equivalently, we found that CRP is associated to a combinatorial logic of global and specific metabolic signals, in coordination with low regulon-size TFs.

Conclusions

What type of questions should we ask in addressing the causes of the emergence of network motifs? Here, we initially focus on two measures of functional specificity of *E. coli*'s TFs based on their corresponding in/out network degree [1,22]. While this association is surely very coarse, it helps us nevertheless to identify two patterns linked to the network simplest motif, i.e., autoregulation. First, TFs with large regulons at the top of the network hierarchy are mostly autoregulated (Fig. 1.C), even though there is a small incidence of this feature in TFs of this layer. This should not be necessarily a surprise, since such global TFs at the top of the hierarchy can elicit a considerable change in bacterial physiology [25,26]. In such scenario, autoregulation not only contributes to the precise integration of environmental states, but can also avoid noisy fluctuations of TF expression [5]. For instance, the autoregulatory circuit of the *crp* gene, one of the TFs at the top of the network hierarchy, plays a major role in CRP signal integration [46], while LexA autoregulation (another top global TF) prevents false (noisy) triggering of the SOS response in *E. coli*, due to transient fluctuations in the inducing signal [14].

A second, and more difficult pattern to interpret, is found in TFs with external transcriptional regulation. In this case, the distribution of autoregulation appears independent of response specificity, i.e., regulon size (Fig. 1.D). Notably, when we quantify the linkage between these TFs and the assembly of more complex motifs (specifically, its role as *Y* element of a FFL or FFLness, Fig. 2), this reveals a strong dependence between response and motif appearance. TFs enabling specific responses (small regulon size) tend to establish relatively more FFLs with their regulated genes than those inducing less specific responses (large regulon size; the decay of FFLness with regulon size is generally observed, i.e., even when specificity classes are not explicitly considered, data not shown).

That low regulon-size TFs –with upstream transcriptional regulation– tend to constitute FFLs with most of their regulon could be ascribed to several neutral constraints, and we analyze the two *a priori* more direct ones, i.e., genome architecture and homology of the constituents of the FFL. Are these factors fully explaining this tendency? The answer appears to be no. Indeed, a careful functional analysis of this family of FFLs highlights a hierarchical logic mostly linked to catabolite repression (Fig. 3.A). This logic is also found in autoregulated polycistrons which points at stronger selective forces acting on this type of regulation than on the specific genetic implementation. Which forces ultimately determine either architecture is hard to tell. In summary, the previous reasoning presents this class of aggregated FFLs as isolated working units beyond those arguments relying uniquely on statistical overrepresentation (see below).

The analysis of the linkage between TF response specificity and FFL emergence provides another relevant pattern. This is the drastic decay in the tendency to assemble FFLs when the *Y* element is a hub regulator. Since the neutral likelihood of establishing a FFL scales with regulon size, we ask to what extend those FFLs with *Y*s being hubs were mostly nonadaptive.

To answer this, and considering that most *X*s in the extant network are hubs, we contrast the coregulation between hubs observed in *E. coli* with the averaged value obtained in a null network model. We find that only a small set of coregulations appear significant under this null (Table 1), but exclude one of them – (SoxS, MarA)– as adaptive since it exhibits duplication. Interestingly, the rest of potentially adaptive coregulations lead to a number of FFL and bifan aggregates with remarkable function coordination, e.g., (FliA, FlhDC) related to flagellar control or several bifans associated to IHF, the integrator host factor regulator. Finally, by investigating a complementary strategy to coregulation, i.e., exclusive regulation, and contrasting it again to a network null, we identify a small group of SIMs –mostly stress response systems (Table 2)– that could also be putatively considered as adaptive, e.g., the LexA DNA damage response SIM.

Overall, the results presented here help to better understand the explicit functional signature behind the statistical definition of network motifs in *E. coli*. These motifs were originally recognized as patterns recurrently found in the extant transcriptional network when compared to a degree-preserving random one. This was done by using a summarized statistical score linked to any considered circuit architecture, i.e., counting the number of regulatory patterns of any particular type and comparing it with the null value. This work used the same kind of null-hypothesis model to show that not all constituents of a given motif class are equally unexpected. We also argued that those that appeared adaptive (like the FFLs with low regulon-size *Y*) could be subjected to other selective forces not necessarily linked to the computational tasks associated to the motif. For the rest of motifs appearing neutral, it is difficult to reject them a priori as adaptive units, as some sort of selection to maintain these edges in the network should be at work (e.g., [5]). Moreover, from the global statistical overrepresentation arguments that led to the description of the network motifs, and even assuming that each of the regulatory links has been selected, one cannot deduce that each of the extant motifs is under selection as a functional entity. Thus, this work sharpens the original counting arguments and contributes to the observations that more elaborated neutral models (see, [15,17,19,7,47]) are required to fully understand the adaptive dynamics of biological networks.

Materials and Methods

Network data

We assembled a transcriptional regulatory network (TRN) with data from *Escherichia coli*'s RegulonDB (v.5.6) [2]. In this network, each interaction is given by the operon encoding the transcriptional factor (TF), that encoding the target gene/s, and a directional link (edge) representing the transcriptional regulation, being this positive, negative or dual (two links of unknown sign were also considered). We did not include those interactions based only on microarrays or undocumented experiments. The TRN is constituted by 681 nodes and 1109 edges between different nodes. 135 of the nodes are TFs, including alternative σ-factors. Within the TF nodes, there are 76 which are autoregulated (approximately 56%), with 12 of them showing no further regulation over any other operon (exclusive autoregulators). The TRN is available in our website (http://www.cnb.csic.es/~jpoyatos. Files *operon_names.txt* and *interactions.txt*, including operon list and specific interactions, respectively).

For additional considerations on the assembly of this TRN see text S1 section 1. We examined several features of this network and that assembled in [4], where the concept of network motifs was originally introduced (Tables S6, S7, S8, S9).

Network null model

We used a null model based on [48], i.e., fixing the number/type of incoming and outgoing edges in the random network to those of *E. coli*'s. The randomization protocol exchanges two randomly chosen connections of the extant network, when both edges are of the same interaction type ($A \rightarrow B, C \rightarrow D$ to $A \rightarrow D, C \rightarrow B$). This procedure is repeated twice the number of edges (2×1109) in order to obtain a fully randomized network (two links of unknown sign were considered as dual ones in these randomizations). This null effectively implies that TF binding sites emerge neutrally. Other statistical methods in text S1 section 2.

Supporting Information

Text S1 Additional analysis and appendix.
Found at: doi:10.1371/journal.pone.0003657.s001 (1.82 MB PDF)

Table S1 Classification of the 230 FFLs in the network based on the connectivity of their respective X- and Y -TFs. LC, MC and HC for low-, medium- and highconnectivity classes, respectively. We also distinguished between autoregulated (curved arrow) and non-autoregulated (crossed-curved arrow) TFs, and those belonging to first (1st-L) and lower-layers (low-L). Small numbers denote number of instances in each subgroup (TFs only regulating their own operon are not considered; Y -elements belong to lower layers of the transcriptional network). The use of the "central unit" association implies an alternative classification of FFLs based on the number of *nonadjacent* regulated operons. Following this criterion, *exuR*, *nagBACD* and *malT*, all regulating one adjacent operon and four nonadjacent ones, are considered low connectivity operons. The minor differences introduced by this latter

classification -which is the one used in Fig. 2.A, main text- are enclosed in parentheses.
Found at: doi:10.1371/journal.pone.0003657.s002 (0.01 MB PDF)

Table S2 Relative orientation between upstream/downstream adjacent genes (→) and TRN operons (⇒). Upstream divergent orientation (←⇒) is particularly enriched. Curved arrow, operons encoding an autoregulated TF; crossed-curved arrow, operons encoding a non-autoregulated TF; *low* curved arrow, operons encoding an autoregulated low-connectivity TF; *low* crossed-curved arrow, operons encoding a non-autoregulated low-connectivity TF; *not*(*low* curved arrow), operons encoding a TF of the TRN excluding autoregulated low-connectivity ones.
Found at: doi:10.1371/journal.pone.0003657.s003 (0.01 MB PDF)

Table S3 First-layer AOs. LC, MC and HC for low-, medium- and high-connectivity classes respectively. In LC without adjacent regulation we distinguish the cases of polycistronic and monocistronic AOs. † d, divergent; u, unidirectional. ‡ Regulated second neighbors included. Calculations based only on microarray data enclosed in brackets. Ψ In those cases with adjacent regulation, we showed number of promoters corresponding to the autoregulated and the adjacent operon, respectively.
Found at: doi:10.1371/journal.pone.0003657.s004 (0.01 MB PDF)

Table S4 Lower-layers AOs of low-connectivity class. When there is not adjacent regulation we distinguish the cases of polycistronic and monocistronic AOs. † d, divergent; c, convergent; u, unidirectional. In the *rhaSR* case there is adjacent regulation over both the upstream and downstream neighbors. ‡ Regulated second neighbors included. Calculations based only on microarray data enclosed in brackets. Ψ In those cases with adjacent regulation, we showed number of promoters corresponding to the autoregulated and the adjacent operon, respectively.
Found at: doi:10.1371/journal.pone.0003657.s005 (0.01 MB PDF)

Table S5 Lower-layers AOs of medium- (MC) and high-connectivity (HC) classes. † d, divergent; u, unidirectional. ‡ Regulated second neighbors included. Calculations based only on microarray data enclosed in brackets. Ψ In those cases with adjacent regulation, we showed number of promoters corresponding to the autoregulated and the adjacent operon, respectively. ¶ *cmk-rpsA-ihfB* and *thrS-infC-rpmI-rplT-pheMST-ihfA*, encoding the two components of the transcription factor IHF, counted as a single node in the network (see the first section of text S1).
Found at: doi:10.1371/journal.pone.0003657.s006 (0.01 MB PDF)

Table S6 General features of SO and CP networks. Curved arrow, operons encoding an autoregulated TF (autoregulated operons); crossed-curved arrow, operons encoding a non-autoregulated TF. Operons encoding a TF that only regulates its own operon in parentheses.
Found at: doi:10.1371/journal.pone.0003657.s007 (0.01 MB PDF)

Table S7 Comparison between autoregulated operons in SO and CP networks. An autoregulated operon in the CP network can be autoregulated (curved arrow), non-autoregulated (crossed-curved arrow) or absent (Abs) in the SO network, and conversely. We specified those operons located in first and lower network layers. Operons appearing in the network only as target operons in parentheses.
Found at: doi:10.1371/journal.pone.0003657.s008 (0.01 MB PDF)

Table S8 Coherent and incoherent FFLs in SO and CP networks (as defined in ref. [2], text S1). Coh: coherent FFLs; Inc: incoherent FFLs, Other: FFLs with at least one dual-type interaction (see also note 3 in text S1).
Found at: doi:10.1371/journal.pone.0003657.s009 (0.00 MB PDF)

Table S9 Distribution of operons per layer in SO and CP networks. We showed explicitely the distribution of autoregulated (curved arrow) and non-autoregulated TF (crossed-curved arrow). † The two components of the *marRAB-rob* loop are considered to be located both in the 6th layer.
Found at: doi:10.1371/journal.pone.0003657.s010 (0.01 MB PDF)

Table S10 Characterization of low-connectivity Y -TFs establishing FFLs with at least one nadZ. First and second columns: Y and X TFs -homolog pairs in bold (two-component systems are also shown). Third and fourth columns: functional characterization of proteins in the central unit and corresponding nadZs labeled with numbers. This also shows the homology relationship -highlighted by same color- between genes in nadZs and those in the associated central unit. Abbreviations: TF, transcriptional factor; 2c, two-component system; E, Enzyme; T, transporter; PTAE, periplasmic transportassociated enzyme; U, uncharacterized protein; NP, near pathway, products acting in regions of the metabolic pathway near those of the central unit; RP: redundant pathway, including proteins which constitute multienzymatic complexes with those encoded in the central unit; P: pathway, sometimes there is no pathway encoded in the central unit, but in the nadZs. See Appendix in text S1 for further details.
Found at: doi:10.1371/journal.pone.0003657.s011 (0.00 MB PDF)

Figure S1 Regulatory links associated to lower-layers operons encoding a low-connectivity autoregulated TF (1≤out-degree<5). We showed incoming and outgoing regulations and also those additional ones to describe FFLs (X-Z interactions). Edges color code: blue, activation; red, repression; gray, dual regulation. Z-operons filling color code: black, Z- and Y -operon are adjacent; gray, Z and Y are second neighbors; white, Z and Y are not adjacent. Dashed lines denote links where the TF encoded in the autoregulated operon is not affected by the regulation. This particularly applies to the regulation of *pdhR-aceEFlpdA* by *arcA*, and leads to the constitution of two pseudo-FFLs. Abbreviations: <*rpo*>, *nlpD-rpoS*; <*hyp*>, *hypABCDE-fhlA*; <*hyc*>, *hycABCDEF-GHI*; <*hyf*>, *hyfABCDEFGHIJR-focB*; <*rpoN*>, *lptB-rpoN-yhbH-ptsN-yhbJ-npr*; <*ihf*>, *cmkrpsA-ihfB*; <*csiD*>, *csiD-ygaF-gabDTP*; <*bae*>, *mdtABCD-baeSR*; <*pdhR*>, *pdhRaceEF- lpdA*; <*srl*>, *srlAEBD-gutM-srlR-gutQ*; <*tdcA*>, *tdcABCDEFG*. Averaged FFLness: <*F*> = 0.77.
Found at: doi:10.1371/journal.pone.0003657.s012 (0.02 MB PDF)

Figure S2 Regulatory links associated to lower-layers operons encoding a low connectivity non-autoregulated TF (out-degree<5). Abbreviations: <*ompR*>, *ompR-envZ*; <*yiaK*>, *yiaKLMNO-lyxK-sgbHUE*, rest of abbreviations as before. Color coding as in Figure S1. <*F*> = 0.46.
Found at: doi:10.1371/journal.pone.0003657.s013 (0.02 MB PDF)

Figure S3 Regulatory links associated to lower-layers operons encoding a medium connectivity autoregulated TF (5≤out-degree<10). In the alternative classification of TFs based on the number of nonadjacent regulated operons *nagBACD* is considered a low-connectivity operon. Maximal FFLness of *rcsA*, *glnALG* and *cytR* corresponds to pairs (X,Y) in which the action of one TF totally relies on the presence of its partner (RcsA on RcsB, RpoN on NtrC -encoded in glnG- and CytR on CRP). Abbreviations: <*mraZ*>, *mraZW-ftsLI-murEF-mraYmurD- ftsW-murGC-ddlB-ftsQAZ*; <*wza*>, *wza-wzb-wzc-wcaAB*; <*mutY*>, *mutYyggX- mltC-nupG*, rest of abbreviations as before. Color coding as in Figure S1. <*F*> = 0.66.
Found at: doi:10.1371/journal.pone.0003657.s014 (0.02 MB PDF)

Figure S4 Regulatory links associated to lower-layers operons encoding a medium connectivity non-autoregulated TF (5≤out-degree<10). In the alternative classification of TFs based on the

number of nonadjacent regulated operons malT is considered a low-connectivity operon. The type of transcriptional interaction between *cmk-rpsA-ihfB* and *flhDC* is not known (in black). Abbreviations: <*malK*>, *malK-lamB-malM*, rest of abbreviations as before. Color coding as in Figure S1. <*F*> = 0.39.
Found at: doi:10.1371/journal.pone.0003657.s015 (0.01 MB PDF)

Acknowledgments

We thank Uri Alon for comments on an earlier version of the manuscript.

Author Contributions

Conceived and designed the experiments: FMC JFP. Performed the experiments: FMC JFP. Analyzed the data: FMC JFP. Contributed reagents/materials/analysis tools: FMC JFP. Wrote the paper: FMC JFP.

References

1. Oltvai Z, Barabási AL (2004) Network Biology: understanding the cell's functional organization. Nat Rev Genet 5: 101–113.
2. Salgado H, et al. (2006) RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res 34: D394–D397.
3. Thieffry D, Huerta AM, Pérez-Rueda E, Collado-Vides J (1998) From specific gene regulation to genomic networks: A global analysis of transcriptional regulation in *Escherichia coli*. BioEssays 5: 433–440.
4. Shen-Orr SS, Milo R, Mangan S, Alon U (2002) Network motifs in the transcriptional regulation network of *Escherichia coli*. Nat Genet 31: 64–68.
5. Alon U (2007) Network motifs: theory and experimental approaches. Nat Rev Genet 8: 450–461.
6. Prill RJ, Iglesias PA, Levchenko A (2005) Dynamic properties of network motifs contribute to biological network organization. PLoS Biology 3: e343.
7. Milo R, et al. (2004) Superfamilies of evolved and designed networks. Science 303: 1538–1542.
8. Ishii T, Yoshida K, Terai G, Fujita Y, Nakai K (2001) DBTBS: a database of Bacillus subtilis promoters and transcription factors. Nucleic Acids Res 29: 278–280.
9. Lee TI, et al. (2002) Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 298: 799–804.
10. Mangan S, Zaslaver A, Alon U (2003) The coherent feedforward loop serves as a sign-sensitive delay element in transcription networks. J Mol Biol 334: 197–204.
11. Kalir S, Alon U (2004) Using a quantitative blueprint to reprogram the dynamics of the flagella gene network. Cell 117: 713–20.
12. Kalir S, Mangan S, Alon U (2005) A coherent feed-forward loop with a SUM input function prolongs flagella expression in *Escherichia coli*. Mol Syst Biol 1: 2005.0006.
13. Mangan S, Itzkovitz S, Zaslaver A, Alon U (2006) The incoherent feed-forward loop accelerates the response-time of the gal system of *Escherichia coli*. J Mol Biol 356: 1073–1081.
14. Camas FM, Blázquez J, Poyatos JF (2006) Autogenous and nonautogenous control of response in a genetic network. Proc Natl Acad Sci U S A 103: 12718–12723.
15. Lynch M (2007) The evolution of genetic networks by nonadaptive processes. Nat Rev Genet 8: 803–813.
16. Banzhaf W, Kuo PD (2004) Network motifs in natural and artificial transcriptional regulatory networks. J Biol Phys Chem 4: 85–92.
17. Cordero OX, Hogeweg P (2006) Feed-forward loop circuits as a side effect of genome evolution. Mol Biol Evol 10: 1931–1936.
18. Ward JJ, Thornton JM (2007) Evolutionary models for formation of network motifs and modularity in the Saccharomyces transcription factor network. PLoS Comput Biol 3: 1993–2002.
19. Artzy-Randrup Y, Fleishman SJ, Ben-Tal N, Stone L (2004) Comment on "Network motifs: simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". Science 305: 1107.
20. Madan BM, Teichmann SA, Aravind L (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol 358: 614–633.
21. Mazurie A, Bottani S, Vergassola M (2005) An evolutionary and functional assessment of regulatory network motifs. Genome Biol 6: R35.
22. Alon U (2007) An Introduction to Systems Biology. London: Chapman and Hall/CRC.
23. Anderson JC, Voight CA, Arkin AP (2007) Environmental signal integration by a modular AND gate. Mol Syst Biol 3: 133.
24. Ma HW, Buer J, Zeng AP (2004) Hierarchical structure and modules in the *Escherichia coli* transcriptional regulatory network revealed by a new top-down approach. BMC Bioinformatics 5: 199.
25. Balázsi G, Barabási A-L, Oltvai ZN (2005) Topological units of environmental signal processing in the transcriptional regulatory network of *Escherichia coli*. Proc Natl Acad Sci U S A 102: 7841–7846.
26. Yu H, Gerstein M (2006) Genomic analysis of the hierarchical structure of regulatory networks. Proc Natl Acad Sci U S A 103: 14724–14731.
27. Cosentino Lagomarsino M, Jona P, Bassetti B, Isambert H (2007) Hierarchy and feedback in the evolution of the *Escherichia coli* transcription network. Proc Natl Acad Sci U S A 104: 5516–5520.
28. Warren PB, ten Wolde PR (2004) Statistical analysis of the spatial distribution of operons in the transcriptional regulation network of *Escherichia coli*. J Mol Biol 342: 1379–390.
29. Hershberg R, Yeger-Lotem E, Margalit H (2005) Chromosomal organization is shaped by the transcription regulation network. Trends Genet 21: 138–142.
30. Korbel JO, Jensen LJ, von Mering C, Bork P (2004) Analysis of genomic context: Prediction of functional associations from conserved bidirectionally transcribed gene pairs. Nat Biotechnol 22: 911–917.
31. Kolesov G, Wunderlich Z, Laikova ON, Gelfand MS, Mirny LA (2007) How gene order is influenced by the biophysics of transcription regulation. Proc Natl Acad Sci U S A 104: 13948–13953.
32. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA (2004) Structure and evolution of transcriptional regulatory networks. Curr Opin Struct Biol 14: 283–291.
33. Teichmann SA, Babu MM (2004) Gene regulatory network growth by duplication. Nat Genet 36: 492–496.
34. Pál C, Papp B, Lercher MJ (2005) Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. Nat Genet 37: 1372–1375.
35. Lozada-Chávez I, Janga SC, Collado-Vides J (2006) Bacterial regulatory networks are extremely flexible in evolution. Nucleic Acids Res 34: 3434–3445.
36. Price MN, Arkin AP, Alm EJ (2006) The life-cycle of operons. PLoS Genet 2: e96.
37. Price MN, Dehal PS, Arkin AP (2008) Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*. Genome Biology 9: R4.
38. Lercher MJ, Pal C (2007) Integration of horizontally transferred genes into regulatory interaction networks takes many million years. Mol Biol Evol; 10.1093/molbev/msm283.
39. Dobrin R, Beg QK, Barabási A-L, Oltvai ZN (2004) Aggregation of topological motifs in the *Escherichia coli* transcriptional regulatory network. BMC Bioinformatics 5: 10.
40. Kashtan N, Itzkovitz S, Milo R, Alon U (2004) Topological generalizations of network motifs. Phys Rev E 70: 031909.
41. Janga SC, Salgado H, Martınez-Antonio A, Collado-Vides J (2007) Coordination logic of the sensing machinery in the transcriptional regulatory network of *Escherichia coli*. Nucleic Acids Res 35: 6963–6972.
42. Missiakas D, Raina S (1998) The extracytoplasmic function sigma factors: role and regulation. Mol Microbiol 28: 1059–1066.
43. Goosen N, van de Putte P (1995) The regulation of transcription initiation by integration host factor. Mol Microbiol 16: 1–7.
44. Bradley MD, Beach MB, de Koning AP, Pratt TS, Osuna R (2007) Effects of Fis on *Escherichia coli* gene expression during different growth stages. Microbiology 153: 2922–2940.
45. Hantke K (1981) Regulation of ferric iron transport in *Escherichia coli* K12: isolation of a constitutive mutant. Mol Gen Genet 182: 288–292.
46. Ishizuka H, Hanamura A, Inada T, Aiba H (1994) Mechanism of the down-regulation of cAMP receptor protein by glucose in *Escherichia coli*: role of autoregulation of the crp gene. EMBO J 13: 3077–3082.
47. Milo R, Itzkovitz S, Kashtan N, Levitt R, Alon U (2004) Response to comment on "Network motifs: Simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". Science 305: 1107.
48. Maslov S, Sneppen K (2002) Specificity and stability in topology of protein networks. Science 296: 910–913.

Supplement

What determines the assembly of transcriptional network motifs in *Escherichia coli*?

Francisco M. Camas and Juan F. Poyatos
Spanish National Biotechnology Centre, Consejo Superior de Investigaciones Científicas (CSIC), 28049 Madrid, Spain.

1 Transcriptional network

This section includes some specifications on the assembled transcriptional regulatory network (TRN) and the quantification of its main attributes (e.g., measurement of auto-regulation in the network):

Heterodimers as single nodes. The IhfA and IhfB constituents of the heterodimer regulator IHF are encoded in two different operons, *thrS-infC-rpmI-rplT-pheMST-ihfA* and *cmk-rpsA-ihfB*, respectively. Because of their similar genomic architecture and regulation (IhfA and IhfB always work as heterodimer and are both under the same regulation), these operons are represented by a single node in the TRN. Similar reasoning applies to the HupA and HupB transcription factors (TFs), components of the heterodimer HU.

Heterodimers as two nodes. The heterodimer RcsAB, whose corresponding operons encode RcsA and RcsB, does not show the previous behavior. RcsB works independently as a homodimer activator (member of the 2-component system RcsC/RcsB). Moreover, RcsAB regulates *rcsA* but not *rcsB*. We thus considered *rcsA* as an autoregulated operon (AO), with the assistance of the protein RcsB, and the operons encoding RcsA and RcsB as two nodes in the TRN.

gntRKU **operon.** We interpreted *gntRKU* as two separated operons (*gntR* and *gntKU*). This prevents the pseudo-autoregulation of *gntKU* by a constitutive GntR, in which GntR would not regulate itself. This also impedes IdnR regulation over *gntKU* (but not over *gntR*) to establish a "pseudo-loop" (or non-dynamical loop) between *gntRKU* and *idnDOTR*.

Transcriptional feedback loops. A recent study documented four transcriptional feedback loops –with more than one component– in *Escherichia coli*'s

transcriptional network [1]. In our TRN only one of these loops remains. Why is this so? One missing loop is the previously mentioned case of non-dynamical loop constituted by *gntRKU*/*idnDOTR*. The other two missing loops appeared when regulations based only on microarray data were considered, and thus they did not occur in our TRN. The only loop that we did find was that constituted by the *marRAB* and *rob* operon pair (Figure S 4).

Feed-forward loop motifs. We identified 232 feed-forward loops (FFLs) in the TRN (Table S1, and Figure 3.A, main text) [1]. In this list there are two instances that should be considered as "pseudo-FFLs". By this we refer to those motifs in which the gene encoding the Y-TF is not part of any transcription unit (TU) regulated by the X-TF (recall that in a FFL $X \rightarrow Y$, $Y \rightarrow Z$, and $X \rightarrow Z$). In both cases, although *arcA* and *pdhR-aceEF-lpdA* are annotated as the X- and Y-elements, respectively, ArcA only regulates the TU constituted by the *lpdA* gene, which is not including the TF acting as putative Y-element of these FFLs, i.e., PdhR [2].

Comparison between Shen-Orr *et al.*, and Camas and Poyatos transcriptional networks. We examined several features of our TRN (CP network) and that assembled in [3] (SO network), where the concept of network motifs was originally introduced. This includes comparisons on 1) network main properties (Table S6), 2) number of AOs (Table S7), 3) FFLs (Table S8), and 4) distribution of operons in the network multilayered structure (Table S9).

2 Main statistical procedures

Autoregulation. We asked two questions related to the distribution of autoregulation in the TRN. First, we examined the distribution of the 64 autoregulated TFs (we did not consider the exclusive autoregulators) with respect to TF sensing specificity. We used a permutation test in which we maintained the number of TFs with and without upstream regulation but randomized the location of the autoregulated TFs. We then measured the number of

[1] Figures S1-S4 show the incoming and outgoing regulations of low/medium regulon-size Y-operons. We showed also the additional links that constitute FFLs.

[2] A list with the 232 FFLs can be found in our website, http://www.cnb.csic.es/∼jpoyatos. The file *all_FFL.txt* contains the X-, Y- and Z-operons listed in the first three columns. We added a fourth column with the FFL class as defined in [2].

autoregulations located among those TFs without external control (i.e. the first layer of the TRN) and compared to the observed value. The presence of autoregulated TFs in this group is smaller than expected ($p = 0.03$, 10000 randomizations). Second, we analyzed how autoregulation correlated with response specificity, i.e., regulon size. We used a permutation test in which we randomized de location of the autoregulations preserving group size of each specificity class (Fig. 1.B, main text). We repeated this protocol in the subsets of TFs with and without upstream control. Thus, only autoregulations inside each subset were randomized (Fig. 1.C-D). Only hubs without upstream control showed a significant enrichment of autoregulated TFs ($p = 0.0086$, 10000 permutations). Alternatively, low regulon-size TFs lacking upstream regulation exhibited a significant low rate of autoregulation ($p = 0.02$, 10000 permutations).

FFLness. We introduced in the main text FFLness ($\mathcal{F}$) as a measure applicable only to TFs with upstream control and regulating ≥ 1 operon(s) –not including autoregulation. For any of these TFs, $\mathcal{F}$ is the ratio of the number of the FFLs in which the TF acts as Y-element, and the maximum number of FFLs that could be potentially assembled with the number of TFs regulating Y (n_{in}) and its regulon size, n_{out} (Fig. 2.A, main text). To examine the significance of the observed measure, we compared with the mean $\mathcal{F}$ obtained in a network null model, controlling for specificity class. Note that i) FFLness is a normalized magnitude that highlights the statistical relevance of the constituted FFLs, i.e., a few FFLs could be easily assembled in a random way by TFs with large regulons, ii) FFLness is almost independent of regulon-size in the null model (Fig. 2.B-D, main text, continuous gray lines), which shows how this magnitude does not exhibit specificity-dependent biases, and iii) the small value of random $\mathcal{F}$ reflects the small number of FFLs that are constituted on average in the random networks (~ 100 *vs.* 232 in the extant network).

When considering the total set of TFs with upstream regulation (Fig 2.D), and comparing with random networks, we found a significantly high $\mathcal{F}$ for all regulon-size classes (low and medium class TFs, $p < 10^{-4}$; hubs, $p = 8 \times 10^{-4}$). FFLness also significantly decayed with regulon size ($p = 0.004$, comparing FFLness of low and high regulon-size TFs, Wilkoxon rank sum test). The use of the alternative class definition discussed in Table S1 showed similar qualitative results. The specific regulatory interactions associated to the computation of $\mathcal{F}$ for low and medium regulon-size TFs

are plotted in Figures S1-S4.

We applied the same protocol to the subsets of autoregulated/non autoregulated TFs (Fig 2.B-C, main text). We found the same qualitative pattern as before. Although the slope of FFLness decay is larger for autoregulated TFs, we did not found a significant difference (see main text for numerical results and their comparison with those considering adjacent regulation).

Significant coregulations by hubs. We counted how many coregulations were established on average by each possible pair of hubs (23 hubs, 253 pairs) in 10000 randomized networks and compared it with those of the extant network. We obtained in this way a set of 253 unadjusted p-values that were corrected for multiple testing as described next.

FDR. We controlled the False Discovery Rate in situations of multiple testing, i.e., when several p-values are calculated simultaneously. We used the following procedure [4]: let $p_1 \leq p_2 \leq \ldots \leq p_m$ be a set of (ordered) unadjusted p-values, the corresponding adjusted p-values are computed as $\tilde{p}_j = min_{k=1,\ldots,m} \left\{ min \left(\frac{m}{k} \, p_k, 1 \right) \right\}, j = 1, \ldots, m$.

Significant SIMs. SIM motifs correspond to TFs exclusively regulating ≥ 3 operons (under the same interaction type). There are 36 TFs that could act potentially as master regulators of positive SIMs in random networks (i.e., each of them regulating $\geq$3 operons –exclusively or not– with positive sign) and 35 TFs as master regulators of negative ones. For each of these TFs we counted how many operons they regulated in a exclusive way in a set of 10000 randomizations and compared this random score with the one observed in *E.coli* (p-values of positive and negative SIMs were adjusted independently).

3 Genomic features of the autoregulated operons

Orientation of genes adjacent to the TRN operons. Divergent architectures can promote the coregulation of the flanking operons through the shared regulatory region (Fig. 2.E, main text). In particular, when the regulation is exerted by a TF encoded in one of these operons, neighbor regulation and autoregulation are readily associated [5]- [8]. To complement the discussions on this issue in the main text, we asked to what extent this divergent architecture has been selected.

We analyzed the relative orientation of the upstream adjacent gene to each of the 681 operons part of the network (Table S2). Note that such adjacent genes could not be constituents of the TRN. We compared this behavior with a null score (randomizing operon orientations in *E.coli*'s genome, 10000 times, while keeping fixed the number of operons encoded in each strand).

Divergent orientations are particularly observed (Table S2). This bias is stronger in the subset of autoregulated operons and was not observed among non-autoregulated ones. We analyzed this significant signal and observed that it was only found (and further enhanced) in the subset of autoregulated operons without upstream control (Table S2, $\not\rightarrow\circlearrowright$). Note that the orientation of adjacent and *downstream* genes did not show any special bias.

Operon structure. We examined the polycistronic/monocistronic architecture of those AOs that, being part of the low regulon-size class, did not regulate an adjacent operon. While there is no particular bias to either design in those AOs without upstream regulation (3 monocistrons + 3 polycistrons, Table S3), polycistronic AOs are considerable enriched in those under this external regulatory control (3 monocistrons + 12 polycistrons, Table S4) [3] . Thus, autoregulated TFs with low regulon-size and upstream regulation are linked both to the polycistronic design and to the assembly of FFLs. These two architectures have in common a dual logic (global TF + specific TF) acting over a set of genes (Figure 3.B-C, main text).

4 Hierarchical FFLs

The "central unit" was defined in the main text as the set constituted by the operon encoding the TF acting as Y and, when applicable, by those of its Z-operons adjacently located (which included adjacent but also second neighbors)[4]. This definition applies to all low regulon-size TFs with upstream regulation (34 operons, Figures S1-S2) and two additional operons (*nagBACD* and *malT*) –both regulating one adjacent operon and four nonadjacent ones, see comments in Table S1. 28 operons of this set are involved, as Y-elements, in the assembly of 74 FFLs (Fig. 3.A, main text). In addition, 53 different operons act as Z-elements of these FFLs.

[3] Among polycistrons associated to TFs with upstream control those with low regulon size are on average the simplest in terms of TUs, even when they are large (Tables S4-S5).

[4] An example of such central unit is the pair of divergent operons plotted in Fig. 3.B, main text.

Approximately half of the previous Y-operons (15/28) regulate at least one non-adjacent Z-operon (second neighbors excluded, see Table S10 and the Appendix of this supplement). There exist 30 of such nonadjacent *Z*s (nad*Z*s), involving 28 different operons (with two cases of shared nad*Z*s: *galETKM* and *manXYZ*, acting as Z-operons of two different Y-operons). Finally, note that to identify homology, we compared amino-acid sequences by Blast with an E-value threshold of 10^{-10} (other threshold values did not change qualitatively our results).

Central unit - nad*Z*s homology. We searched for those nad*Z*s that encode at least one gene homolog to those of the central unit. We obtained 7 out of 30 nad*Z*s with such relationship (Table S10 and Appendix). This number is bigger than expected by chance ($p < 0.0001$ by randomly reassigning 10000 times the set of all nad*Z*s to the set of central units with the restrictions that i) the number of nad*Z*s regulated by the Y-operon of each central unit is fixed, ii) an operon is never assigned to itself [5], and iii) an operon is never assigned twice to the same block because of the mentioned shared *Z*s).

X-Y homology. There are 15 Y-operons regulating nad*Z*s which constitute 42 different (X,Y) pairs with their respective X-operons. We analyzed the homology between genes encoding the X and Y TFs, respectively. We found 6 cases of homolog pairs (Table S10 and Appendix), larger than expected by chance ($p = 0.0003$, by permuting 10000 times TFs and controlling the cases where an operon is paired with itself).[6]

FFLs without homologies. About two thirds (25/40) of the hierarchical FFLs constituted with nad*Z*s cannot be explained by homology-based models (Figure 3.A, main text). We observed that these nad*Z*s are enriched by operons only encoding transport related proteins, and that they are under the control of CRP. These transporters are functionally related to those transporters encoded in the corresponding central unit, yet they are not homologs. Is the transporter located in the central operon, and thus physically linked to the TF, anyway different to those placed nonadjacently? Homologies across transporters associated to different FFLs groups –a given central unit and its associated nad*Z*s– allowed us to compare aspects of function and genomic location. Examples of these homologies are the MFS-symporters or

[5] This could be possible because in the extant network the operon *malT* is both a Y-operon –with four different *Z*s– and a Z-operon of the Y-operon *dgsA*.

[6] *gadW* is found in both X and Y roles.

ABC-transporters of arabinose and galactose (Fig. 4, main text), and also the glucose and the (very related) N-acetyl-D-glucosamine PTS uptake systems. We found equivalent functions encoded in adjacent or nonadjacent locations. We reported in the main text the comparison between the MFS- and ABC-transporters in the arabinose and galactose systems. Additionally, unlike the glucose uptake system located in nad*Z*s, one of the specific components of a N-acetyl-D-glucosamine PTS transporter is encoded in the central unit (see Appendix).

Hierarchical FFLs *vs.* **polycistronic strategies.** We proposed in the main text how an adaptive model based on the establishment of a hierarchical logic on a small set of genes acts as a unifying determinant leading to the occurrence of both hierarchical FFLs and low regulon-size polycistrons with upstream control (Figs. 3.B-C –main text– and Tables S3-S4).

What aspects could influence the presence of either control strategy in a given context? Reasons for the separation in different operons of coregulated genes than act together in a metabolic pathway has been discussed [9]. In brief, this separation allows differential regulation of each operon (enabling temporal programs of gene expression). A polycistronic architecture might not be considered, in this sense, an optimal solution as it could induce the production of some proteins –encoded in the polycistron– before needed. However, this latter strategy can favor the transference of the encoded enzymatic tools across species by horizontal gene transfer (HGT). Neighbor regulation appears in this context as an intermediate solution, combining differential control and capability for successful lateral transfer [7] . Indeed, a large frequency of these events have been recently reported for neighbor regulators [10].

A prediction of the differential expression model [9] is that genes are arranged such that those encoded on the same operon do not skip functional steps in the pathway. This is precisely what we found for genes distributed among the operons in the central unit and the nad*Z*s (see Appendix). Note however that this result could also be due to the mechanisms explaining how bacterial metabolic networks grow, i.e., by HGT uptake of genes encoding products involved in peripheral reactions [11]. This correlates with

[7]The architecture of divergently transcribed operons also reduces the cost of maintenance and replication of an additional promoter region.

the enrichment of nad$Z$s with genes associated to the first steps of peripheral metabolisms [8].

Genome distance between the central unit and the nad$Z$s. For each central unit, we computed the mean distance to its nad$Z$s and then averaged over all units. We then randomized the full set of nad$Z$s and scored distances as before. The average distance of nad$Z$s to the central unit was not particularly small, even when including second neighbors as nonadjacent operons ($p = 0.1$, 10000 randomizations). We also calculated the "across distance" between the coordinates of each central unit and its associated nad$Z$s with respect to the *oriC* region, as chromosomal periodicity of evolutionarily conserved gene pairs has been also recently discussed [12]. This measure did not show any significant pattern either.

Averaged co-conservation of Y- and Z-operons. We considered the phylogenetic conservation of genes involved in the Y/Z operons through 75 species of γ-proteobacteria. Conservation of a particular gene was determined by reciprocal best-hit with an E-value threshold of 10^{-10} (other threshold values did not change qualitatively our results). We quantified co-conservation of each Y-operon/Z-operon by first averaging the Jaccard index [9] of proximity J for all the possible pairs of genes $(y,\ z)/y \in Y,\ z \in Z$. We then determined the average value of J over the set of 30 pairs constituted by the nad$Z$s with their respective Y-operons, and also for the 10 pairs with adjacent Z-operons (ad$Z$s, including here the second neighbors). The average co-conservation of the pairs $\{Y$, all associated Zs –adjacent or not–$\}$ was significantly larger than expected by randomly reassigning the set of Zs ($p < 10^{-3}$, 10000 permutations) [10] . Moreover, the difference on this averaged co-conservation for nad$Z$s (0.40) and ad$Z$s (0.43) was not significant under the permutation of the ad$Z$/nad$Z$ labels (p=0.32, 10000 times).

[8] The most unquestionable cases of non-neutral evolution among hierarchical FFLs are those constituted with nad$Z$s and which could not be explained by homology-based models (25 cases, Fig. 3.A main text). 19 different operons act as nad$Z$s in these FFLs, from which 12 only encode transport related products –also associated to HGT events [11].

[9] This normalized index is a ratio of the number of species in which both genes coexist divided by the total number of species considered. As a reference, the mean value of J for pairs of genes belonging to the same operon is 0.64 (for this set of Y- and Z-operons).

[10] To avoid that the signal of large co-conservation were only caused by the adjacent Zs, we applied the same randomization protocol only in the set of nad$Z$s. We obtained again that the averaged co-conservation of the pairs $\{Y$, nad$Z$s$\}$ was significantly large (p=0.02, 10000 permutations).

Functional characterization. We examined in the Appendix the functional properties of the proteins encoded in the group of 15 low regulon-size Y-operons regulating nad$Z$s (second neighbors excluded, see also Table S10) and all their associated Z-operons, using EcoCyc database [13]. In some cases, the proteins are members of complexes whose additional constituents are not encoded in this group. We nevertheless enclosed this information in parentheses.

We included a simple cartoon showing the location of these proteins in their associated metabolic pathways. We used arrows or ellipses, crossed by arrows, to denote enzymes and transporters, respectively. When a protein is encoded in the central unit, we colored the corresponding symbol in blue. We used red for proteins encoded in nad$Z$s, and gray for proteins encoded in other operons. Some protein complexes required two colors at the same time.

We also described the previously discussed gene homologies, i.e, those between the central unit and nad$Z$s and those between TFs acting as X- and Y-elements of the FFL. Furthermore, we showed for adjacent regulations the relative direction of transcription with respect to that of the Y-operon: (d), divergent; (u), unidirectional (convergent cases were not found). We also indicated when the ad$Z$ is a second neighbor. Abbreviations: Y-op, Y-operon; nad$Z$, nonadjacent Z-operon; ad$Z$, Z-operon adjacent to the Y-operon (including second neighbors).

References

[1] Cosentino M, Jona P, Bassetti B, Isambert H (2007) *Proc Natl Acad Sci USA* 104:5516–5520.

[2] Mangan S, Alon U (2003) *Proc Natl Acad Sci USA* 100:11980–11985.

[3] Shen-Orr SS, Milo R, Mangan S, Alon U (2002) *Nature Genet* 31:64–68.

[4] Benjamini Y, Hochberg Y (1995) *J R Statist Soc B* 57:289–300.

[5] Kolesov G, Wunderlich Z, Laikova ON, Gelfand MS, Mirny LA (2007) *Proc Natl Acad Sci U S A* 104:13948–13953.

[6] Korbel JO, Jensen LJ, von Mering C, Bork P (2004) *Nat Biotechnol* 22:911–917.

[7] Warren PB, ten Wolde PR (2004) *J Mol Biol* 342:1379–390.

[8] Hershberg R, Yeger-Lotem E, Margalit H (2005) *Trends Genet* 21:138–142.

[9] Zaslaver A, Mayo A, Ronen M, Alon U (2006) *Phys Biol* 3:183–9.

[10] Price MN, Dehal PS, Arkin AP (2008) Horizontal gene transfer and the evolution of transcriptional regulation in *Escherichia coli*. *Genome Biology* 9:R4

[11] Pál C, Papp B, Lercher MJ (2005) *Nat Genet* 37:1372–1375.

[12] Wright MA, Kharchenko P, Church GM, Segrè D (2007) *Proc Natl Acad Sci USA* 104:10559–10564.

[13] Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gil M, Karp PD (2005) *Nucleid Acids Res* 33:D334–D337.

			X-**TF**												
			LC				MC				HC				
			$\not\to\circlearrowright$	$\not\to\not\emptyset$	$\to\circlearrowright$	$\to\not\emptyset$	$\not\to\circlearrowright$	$\not\to\not\emptyset$	$\to\circlearrowright$	$\to\not\emptyset$	$\not\to\circlearrowright$	$\not\to\not\emptyset$	$\to\circlearrowright$	$\to\not\emptyset$	
			15(16)	30	21(22)	13(14)	5(4)	5	7(6)	4(3)	7	1	9	6	total
Y-**TF**	LC	$\to\circlearrowright$	0(1)	3	0	1	1(0)	2	1	1	27(31)	0	2	9	47(51)
		$\to\not\emptyset$	0	0	0	0	1	0	0	0	14(18)	0	1	3	19(23)
	MC	$\to\circlearrowright$	0	0	0	1	0	6	0	0	13(9)	0	1	8	29(25)
		$\to\not\emptyset$	0	0	0	0	0	0	0	0	9(5)	0	5	0	14(10)
	HC	$\to\circlearrowright$	1	0	0	0	0	0	1	0	29	0	20	10	61
		$\to\not\emptyset$	0	0	0	0	0	2	0	0	53	0	2	5	62
		total	1(2)	3	0	2	2(1)	10	2	1	145	0	31	35	232

Table S1: Classification of the 232 FFLs in the network based on the regulon size of their respective X- and Y-TFs. LC, MC and HC for low-, medium- and high regulon-size classes, respectively. Subgroups are based on the presence/absence of upstream regulation and autoregulation: $\not\to\circlearrowright$, autoregulated TFs without upstream regulation; $\not\to\not\emptyset$, non-autoregulated TFs without upstream regulation; $\to\circlearrowright$, autoregulated TFs with upstream regulation; $\to\not\emptyset$, non-autoregulated TFs with upstream regulation. Small numbers denote number of instances in each subgroup (TFs only regulating their own operon are not considered; Y-elements have upstream regulation by definition). The use of the "central unit"association implies an alternative classification of FFLs based on the number of *nonadjacent* regulated operons. Following this criterion, *exuR*, *nagBACD* and *malT*, all regulating one adjacent operon and four nonadjacent ones, are considered low regulon-size operons. The minor differences introduced by this latter classification –which is the one used in Fig. 3.A, main text– are enclosed in parentheses.

set	N	$\rightarrow\Rightarrow$	$\leftarrow\Rightarrow$	$\Rightarrow\rightarrow$	$\Rightarrow\leftarrow$	p
TRN	681	43.8	56.2	51.0	49.0	0.0015
$\circlearrowleft$	76	36.8	63.2	51.3	48.7	0.02
$\not\circlearrowleft$	59	47.5	52.5	45.8	54.2	0.39
$\not\rightarrow\circlearrowleft$ *low*	18	16.7	83.3	33.3	66.7	0.004
$\rightarrow\circlearrowleft$ *low*	30	36.7	63.3	50.0	50.0	0.10
$\not\circlearrowleft$ *low*	43	48.8	51.2	46.5	53.5	0.50

Table S2: Relative orientation between upstream/downstream adjacent genes ($\rightarrow$) and TRN operons ($\Rightarrow$). Upstream divergent orientation ($\leftarrow\Rightarrow$) is particularly enriched. $\circlearrowleft$, operons encoding an autoregulated TF; $\not\circlearrowleft$, operons encoding a non-autoregulated TF; $\not\rightarrow\circlearrowleft$ *low*, operons encoding an autoregulated low regulon-size TF without upstream regulation; $\rightarrow\circlearrowleft$ *low*, operons encoding an autoregulated low regulon-size TF with upstream regulation; $\not\circlearrowleft$ *low*, operons encoding a non-autoregulated low regulon-size TF (with or without upstream regulation). p, p-value for enrichment of upstream divergent orientation ($\leftarrow\Rightarrow$).

set		AO	Orientation of adj. regulated operon †	Number of nonadjacent regulated op. ‡	Number of promoters in central unit §
LC	**adjacent regulation**	*acrR*	d	0	1/1
		agaR	d	1	1/1
		cusRS	d	0	1/1
		cynR	d	0	1/1
		evgAS	d	1 [1]	2/1
		gcvA	d	1	1/1
		hcaR	d	0	1/1
		ilvY	d	0	1/1
		mngR	d	0	1/1
		pspF	d	1	3/1
		soxR	d	1	1/1
		torR	d	1 [2]	1/1
	poly.	*ada-alkB*	-	2	2
		emrRAB	-	0	1
		qseBC	-	0 [1]	2
	mono.	*lrhA*	-	2	1
		putA	-	0	1
		trpR	-	4	1
MC		*cysB*	-	6 [1]	1
		exuR	u	4	1/1
		iscRSUA	-	6	1
		tyrR	-	7	1
		phoBR	-	9 [1]	1
HC		*argR*	-	10	2
		cpxRA	d	20	1/1
		crp	d	161 [13]	1/1
		fnr	-	85 [7]	1
		lexA-dinF	-	19 [1]	1
		lrp	-	22 [10]	1
		phoPQ	-	19	2

Table S3: Autoregulated operons without upstream regulation. LC, MC and HC for low-, medium- and high regulon-size classes respectively. In LC without adjacent regulation we distinguish the cases of polycistronic and monocistronic AOs. † d, divergent; u, unidirectional. ‡ Regulated second neighbors included. Calculations based only on microarray data enclosed in brackets. § In those cases with adjacent regulation, we showed number of promoters corresponding to the autoregulated and the adjacent operon, respectively.

set	AO	Orientation of adj. regulated operon †	Number of nonadjacent regulated op. ‡	Number of promoters in central unit §
adjacent regulation	*araC*	d	3	1/1
	betIBA	d	0	1/1
	fecIR	u	0	1/1
	galS	u	2	1/1
	glcC	d	0	1/1
	hypABCDE-fhlA	d	3	2/1
	idnDOTR	d	1	1/1
	malI	d	0	1/1
	melR	d	0	1/1
	metR	d	2 [1]	2/1
	prpR	d	0	1/1
	rhaSR	d,c	0	1/1
	uxuR	u	2	1/1
	xylFGHR	d	0	2/1
	zraSR	d	0	1/1
poly.	*chbBCARFG*	-	0	1
	gadAX	-	1 [9]	2
	hipBA	-	0	1
	hyfABCDEFGHIR-focB	-	0	1
	lctPRD (*lldPRD*)	-	0	2
	mdtABCD-baeSR	-	3	1
	mtlADR	-	0	1
	nikABCDER	-	0	2
	pdhR-aceEF-lpdA	-	2	3
	rbsDACBKR	-	0	1
	srlAEBD-gutM-srlR-gutQ	-	0	2
	tdcABCDEFG	-	0	1
mono.	*dgsA* (*mlc*)	-	4	2
	iclR	-	1	1
	nac	-	4 [2]	1

Table S4: Autoregulated operons with upstream regulation and low regulon size. When there is not adjacent regulation we distinguish the cases of polycistronic and monocistronic AOs. † d, divergent; c, convergent; u, unidirectional. In the *rhaSR* case there is adjacent regulation over both the upstream and downstream neighbors. ‡ Regulated second neighbors included. Calculations based only on microarray data enclosed in brackets. § In those cases with adjacent regulation, we showed number of promoters corresponding to the autoregulated and the adjacent operon, respectively.

set	AO	Orientation of adj. regulated operon †	Number of nonadjacent regulated op. ‡	Number of promoters in central unit §
MC	*cytR*	-	8	1
	dnaAN-recF	-	5	8
	gadE	u	5 [8]	3/1
	glnALG	-	5 [7]	3
	nagBACD	d	4	3/1
	oxyR	-	8 [1]	1
	rcsA	-	6 [1]	1
HC	*dusB-fis*	-	54 [8]	1
	fldA-fur	-	31 [4]	4
	fliAZY	u	15	2/1
	hns	-	20 [21]	1
	marRAB	-	15 [1]	1
	purR	-	15 [2]	1
	rpoE-rseABC	-	51	3
	soxS	-	15 [1]	1
	cmk-rpsA-ihfB ¶	-	56 [7]	4
	thrS-infC-rpmI-rplT-pheMST-ihfA ¶			7

Table S5: Autoregulated operons with upstream regulation and belonging to the medium- (MC) and high regulon-size (HC) classes. † d, divergent; u, unidirectional. ‡ Regulated second neighbors included. Calculations based only on microarray data enclosed in brackets. § In those cases with adjacent regulation, we showed number of promoters corresponding to the autoregulated and the adjacent operon, respectively. ¶ *cmk-rpsA-ihfB* and *thrS-infC-rpmI-rplT-pheMST-ihfA*, encoding the two components of the transcription factor IHF, counted as a single node in the network (see the first section of this supplement).

	SO	CP
nodes	423	681
non-autoregulatory interactions	519	1109
TFs	116	135
$\not\rightarrow$	81	66
$\rightarrow$	35	69
$\circlearrowleft$	59 (10)	76 (12)
$\not\rightarrow\circlearrowleft$	35 (5)	30 (3)
$\rightarrow\circlearrowleft$	24 (5)	46 (9)

Table S6: General features of SO and CP networks. $\not\rightarrow$, TFs without upstream regulation; $\rightarrow$, TFs with upstream regulation; $\circlearrowleft$, autoregulated TFs; $\not\rightarrow\circlearrowleft$, autoregulated TFs without upstream regulation; $\rightarrow\circlearrowleft$, autoregulated TFs with upstream regulation. For autoregulators, we detailed the cases of operons encoding a TF that only regulates its own operon (in parentheses).

SO	CP	cases	SO $\not\rightarrow$	SO $\rightarrow$	CP $\not\rightarrow$	CP $\rightarrow$
$\circlearrowleft$	$\circlearrowleft$	50	29	21	20	30
$\circlearrowleft$	$\not\circlearrowleft$	6	3	3	0	2+(4)
$\circlearrowleft$	Abs	3	3	0	-	-
$\not\circlearrowleft$	$\circlearrowleft$	12	9	2+(1)	3	9
Abs	$\circlearrowleft$	14	-	-	7	7

Table S7: Comparison between autoregulated operons in SO and CP networks. An autoregulated operon in the CP network can be autoregulated ($\circlearrowleft$), non-autoregulated ($\not\circlearrowleft$) or absent (Abs) in the SO network, and conversely. We specified those operons with ($\rightarrow$) and without ($\not\rightarrow$) upstream control. Operons appearing in the network only as target operons in parentheses.

	SO	CP
Coh-1	24	66
Coh-2	2	16
Coh-3	4	6
Coh-4	0	9
Inc-1	5	24
Inc-2	0	8
Inc-3	1	2
Inc-4	0	14
Other	6	87
total	42	232

Table S8: Coherent and incoherent FFLs in SO and CP networks (as defined in [2]). Coh: coherent FFLs; Inc: incoherent FFLs, Other: FFLs with at least one dual-type interaction (see also note 2).

	SO network			**CP network**		
layer	operons	⟳	∅⟳	operons	⟳	∅⟳
1	81	35	46	66	30	33
2	233	17	8	177	20	10
3	87	5	3	113	4	3
4	10	2	0	88	6	1
5	12	0	0	65	7	3
6†				94	6	4
7				49	2	2
8				14	1	0
9				15	0	0

Table S9: Distribution of operons per layer in SO and CP networks. We showed explicitely the distribution of autoregulated (⟳) and non-autoregulated TFs (∅⟳). † The two components of the *marRAB-rob* loop are considered to be located both in the 6th layer.

Y-TF	X-TFs	central-unit products	nonadjacent Z-operons products
AraC	CRP	TF, E	1: T; 2: T; 3: T
Cbl	**CysB**	TF	1: P[E, T]
DcuS-DcuR (2c)	Fnr, NarL	2c, E, T	1: NP[E]
DgsA	CRP	TF	1: TF; 2: T; 3: T; 4: T
GadX	CRP, **GadW**, GadE, RpoS	TF, E	1: RP[E, T]
GalS	CRP, **GalR**	TF, T	1: T; 2: P[E]
GlpR	CRP	TF, E, E	1: RP[E]; 2: T, PTAE; 3: NP[T, E], E
HU	CRP	TF	1: P[E]
FhlA	Fnr, IHF, RpoN	TF, E, E	1: RP[TF, E, T]; 2: RP[E]; 3: RP[E, E]
IdnR	CRP, **GntR**	TF, E, E, T	1: RP[E, T]
MalT	CRP	TF, E	1,2: T, T, U; 3: PTAE
BaeS-BaeR (2c)	**CpxA-CpxR** (2c)	2c, T, T	1: T
NagC	CRP	TF, E, T	1: T; 2: NP[TF, E, T]; 3: T
PdhR	CRP, Fnr, ArcA	TF, E	1: NP[TF, E, T]; 2: NP[E]
UxuR	CRP, **ExuR**	TF, E, T	1: NP[E, T]

Table S10: Characterization of low regulon-size Y-TFs establishing FFLs with at least one nad$Z$. First and second columns: Y and X TFs –homolog pairs in bold (two-component systems are also shown). Third and fourth columns: functional characterization of proteins in the central unit and corresponding nad$Z$s labeled with numbers. This also shows the homology relationship –highlighted by same color– between genes in nad$Z$s and those in the associated central unit. Abbreviations: TF, transcriptional factor; 2c, two-component system; E, Enzyme; T, transporter; PTAE, periplasmic transport-associated enzyme; U, uncharacterized protein; NP, near pathway, products acting in regions of the metabolic pathway near those of the central unit; RP: redundant pathway, including proteins which constitute multienzymatic complexes with those encoded in the central unit; P: pathway, sometimes there is no pathway encoded in the central unit, but in the nad$Z$s. See Appendix for further details.

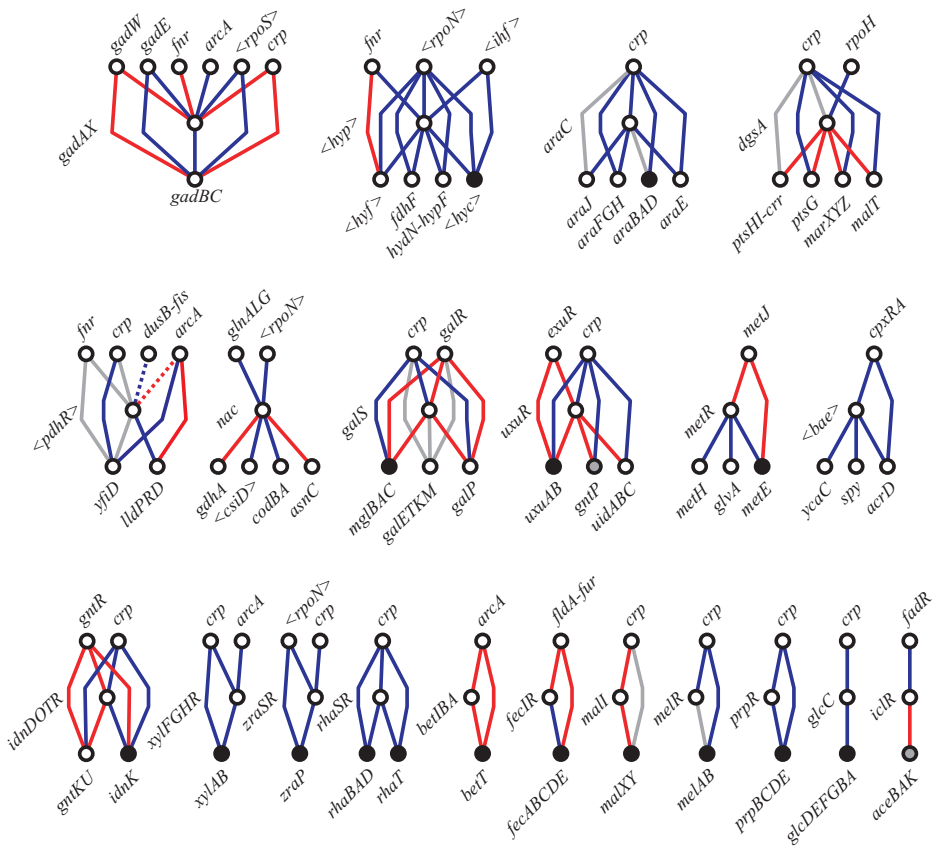


Figure S1: Regulatory links associated to operons with upstream regulation and encoding a low regulon-size autoregulated TF ($1 \leq$out-degree< 5). We showed incoming and outgoing regulations and also those additional ones to describe FFLs (X-Z interactions). Edges color code: blue, activation; red, repression; gray, dual regulation. Z-operons filling color code: black, Z- and Y-operon are adjacent; gray, Z and Y are second neighbors; white, Z and Y are not adjacent. Dashed lines denote links where the TF encoded in the autoregulated operon is not affected by the regulation. This particularly applies to the regulation of *pdhR-aceEF-lpdA* by *arcA*, and leads to the constitution of two pseudo-FFLs. Abbreviations: $\langle$*rpoS*$\rangle$, *nlpD-rpoS*; $\langle$*hyp*$\rangle$, *hypABCDE-fhlA*; $\langle$*hyc*$\rangle$, *hycABCDEFGHI*; $\langle$*hyf*$\rangle$, *hyfABCDEFGHIJR-focB*; $\langle$*rpoN*$\rangle$, *lptB-rpoN-yhbH-ptsN-yhbJ-npr*; $\langle$*ihf* $\rangle$, *cmk-rpsA-ihfB*; $\langle$*csiD*$\rangle$, *csiD-ygaF-gabDTP*; $\langle$*bae*$\rangle$, *mdtABCD-baeSR*; $\langle$*pdhR*$\rangle$, *pdhR-aceEF-lpdA*; $\langle$*srl*$\rangle$, *srlAEBD-gutM-srlR-gutQ*; $\langle$*tdcA*$\rangle$, *tdcABCDEFG*. Averaged FFLness: $\langle \mathcal{F} \rangle = 0.64$ (see Fig. 2.B, main text).

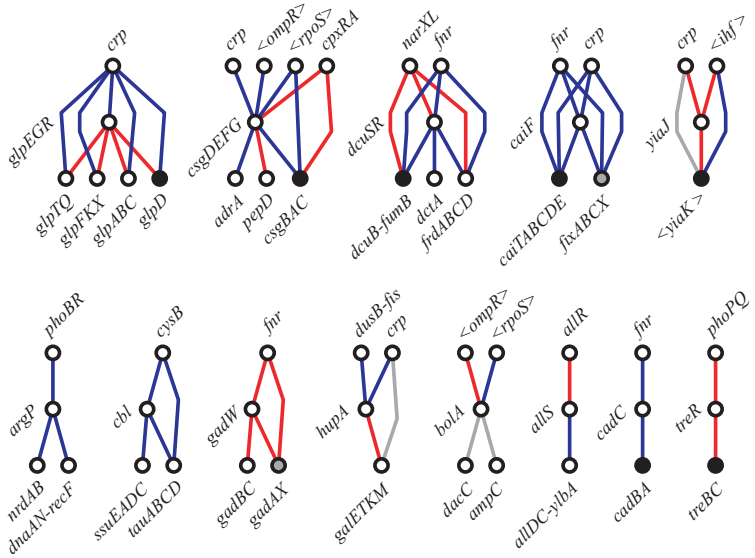


Figure S2: Regulatory links associated to operons with upstream regulation and encoding a low regulon-size non-autoregulated TF ($1 \leq$out-degree< 5). Abbreviations: $\langle$*ompR*$\rangle$, *ompR-envZ*; $\langle$*yiaK*$\rangle$, *yiaKLMNO-lyxK-sgbHUE*, rest of abbreviations as before. Color coding as in Figure S1. $\langle \mathcal{F} \rangle = 0.41$ (see Fig. 2.C, main text).

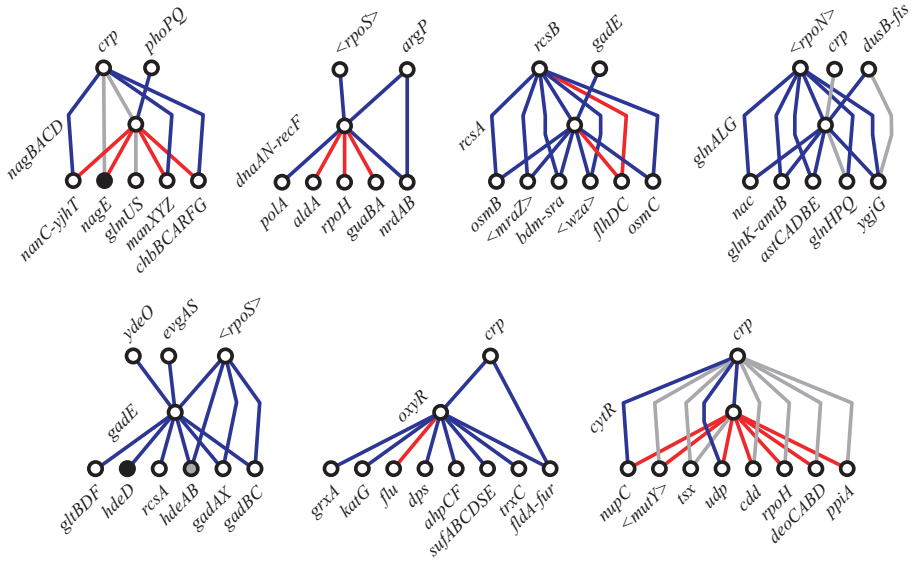


Figure S3: Regulatory links associated to operons with upstream regulation and encoding a medium regulon-size autoregulated TF ($5 \leq$out-degree< 10). In the alternative classification of TFs based on the number of nonadjacent regulated operons *nagBACD* is considered a low regulon-size operon. Maximal FFLness of *rcsA*, *glnALG* and *cytR* corresponds to pairs (X,Y) in which the action of one TF totally relies on the presence of its partner (RcsA on RcsB, RpoN on NtrC –encoded in *glnG*– and CytR on CRP). Abbreviations: $\langle$*mraZ*$\rangle$, *mraZW-ftsLI-murEF-mraY-murD-ftsW-murGC-ddlB-ftsQAZ*; $\langle$*wza*$\rangle$, *wza-wzb-wzc-wcaAB*; $\langle$*mutY*$\rangle$, *mutY-yggX-mltC-nupG*, rest of abbreviations as before. Color coding as in Figure S1. $\langle \mathcal{F} \rangle = 0.38$ (see Fig. 2.B, main text).

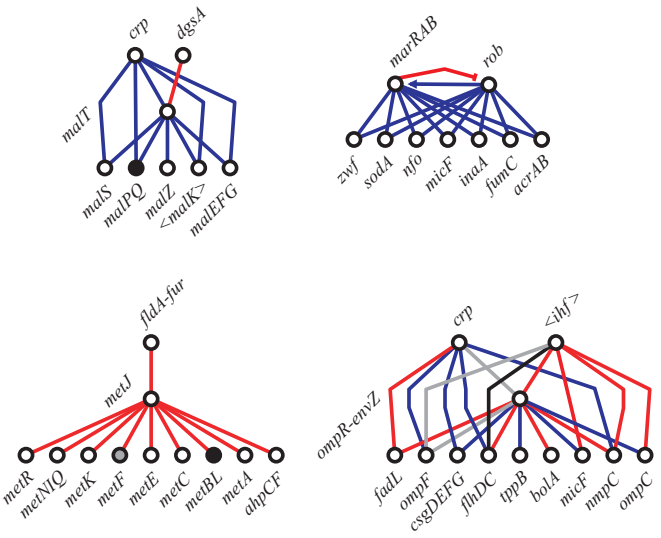


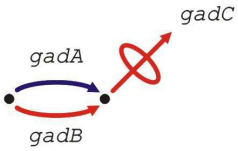
Figure S4: Regulatory links associated to operons with upstream regulation and encoding a medium regulon-size non-autoregulated TF (5≤out-degree<10). In the alternative classification of TFs based on the number of nonadjacent regulated operons *malT* is considered a low regulon-size operon. The type of transcriptional interaction between *cmk-rpsA-ihfB* and *flhDC* is not known (in black) . Abbreviations: ⟨*malK*⟩, *malK-lamB-malM*, rest of abbreviations as before. Color coding as in Figure S1. $\langle \mathcal{F} \rangle = 0.32$ (see Fig. 2.C, main text).

APPENDIX

Autoregulated Y-operon

Y-op: *gadAX*
- *gadA*: enzyme, glutamate dependent acid resistance
- *gadX*: TF

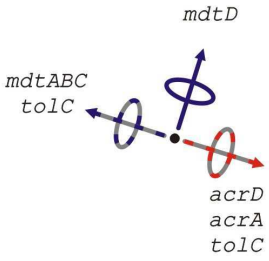
nadZ: *gadBC*
- *gadB*: enzyme, glutamate dependent acid resistance
- *gadC*: APC-transporter (aminobutyrate antiporter)



Notes:
- *gadA* and *gadB* are homologs
- GadX is homolog of the TF encoded in one of its four X-operons, *gadW*. These two operons are second neighbors only separated by the small gene *gadY*

Y-op: *mdtABCD-baeSR*
- *mdtABC* (+ *tolC*): RND-transporter. (drug exporter)
- *mdtD*: MFS-transp. (uncharacterized, drug efflux?)
- *baeSR*: 2-component system

nadZ: *acrD*
- *acrD* (+ *tolC* and *acrA*): RND-transporter (drug exporter)

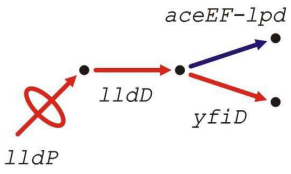


Notes:
- *mdtB*, *mdtC* and *acrD* are homologs
- *baeS* and *baeR* are homologs of the two-component-system genes *cpxA* and *cpxR* respectively; *cpxRA* is the only X-operon for *mdtABCD-baeSR*
- *tolC* encodes the common outer membrane component of several multidrug efflux systems

---

Y-op: *pdhR-aceEF-lpd*
- *pdhR*: TF
- *aceEF-lpd*: pyruvate dehydrogenase

nadZ: *lldPRD*
- *lldP*: LCT-transporter (lactate)
- *lldR*: TF
- *lldD*: lactate dehydrogenase



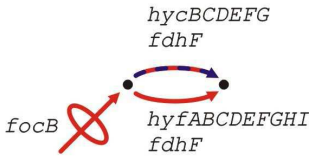
nadZ: *yfiD*
- *yfiD*: alternative stress induced pyruvate-formate lyase

Notes:
- *pdhR* and *lldR* are homologs
- *lldPRD* and *yfiD* are the respective Z-elements of the two pseudo-FFLs (see Fig. S1)
- *lldPRD* is an autoregulated operon

Y-op: *hypABCDE-fhlA*
- *hypABCDE*: proteins for maturation of hydrogenase
- *fhlA*: TF

adZ(d): *hycABCDEFGHI*
- *hycA*: uncharacterized
- *hycBCDEFG*: hydrogenase
- *hycHI*: protein for maturation of hydrogenase

nadZ: *hyfABCDEFGHIJR-focB*
- *hyfABCDEFGHIJ*: hydrogenase (putative)
- *hyfR*: TF
- *focB*: FNT-transporter (formate, putative)



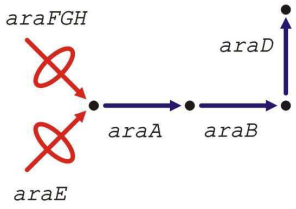
nadZ: *fdhF*
- *fdhF* (+ *hycBCDEFG*): formate-hydrogenlyase complex
- *fdhF* (+ *hyfABCDEFGHIJ*) : formate-hydrogenlyase complex (putative)

nadZ: *hydN-hypF*
- *hydN*: formate-dehydrogenase (putative)
- *hypF*: protein for maturation of hydrogenase

Notes:
- There are multiple homologies between the *hyc* and *hyf* genes
- *fhlA* and *hyfR* are homologs
- *hydN*, *hycB* and *hyfA* are homologs
- *hyfABCDEFGHIJR-focB* is an autoregulated operon

Y-op: *araC*
- *araC*: TF

adZ(d): *araBAD*
- *araBAD*: enzymes in arabinose degradation pathway



nadZ: *araE*
- *araE*: MFS-transporter (arabinose)

nadZ: *araFGH*
- *araFGH*: ABC-transporter (arabinose)

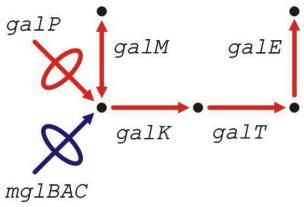
nadZ: *araJ*
- *araJ*: MFS-transporter (uncharacterized, sugar efflux?)

Y-op: *galS*
- *galS*: TF

adZ(u): *mglBAC*
- *mglBAC*: ABC-transporter (galactose)

nadZ: *galP*
- *galP*: MFS-transporter (galactose)

nadZ: *galETKM*
- *galETK*: enzymes for UDP-galactose biosynthesis
- *galM*: galactose-1-epimerase (enzyme that links lactose and galactose metabolisms)



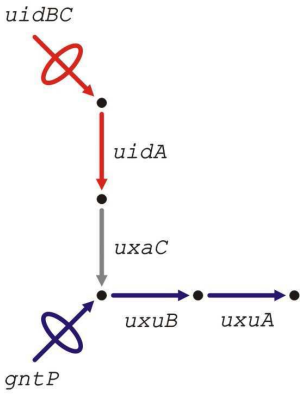
Notes:
- GalS is homolog of the TF encoded in one of its two X-operons, *galR*. The additional X-operon is CRP

Y-op: *uxuR*
- *uxuR*: TF

adZ(u): *uxuAB*
- *uxuAB*: enzymes in fructuronate degradation pathway

adZ(2nd): *gntP*
- *gntP*: GNT-transporter (fructuronate/gluconate)

nadZ: *uidABC*
- *uidA*: enzyme in glucuronide degradation pathway
- *uidB*: GPH-transporter (glucuronide)
- *uidC*: membrane protein associated to *uidB*

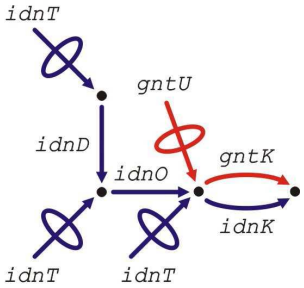


Notes:
- UxuR is homologue of the TF encoded in one of its two X-operons, *exuR*. The additional X-operon is CRP
- *uxuAB* and *gntP* are divergent operons
- *uxaC* is regulated by ExuR. This gene is in the genome neighborhood of *exuR*

Y-op: *idnDOTR*
- *idnDO*: enzymes in idonate degradation pathway
- *indT*: GNT-transporter (idonate/gluconate)
- *idnR*: TF

adZ(d): *idnK*
- *idnK*: enzyme in idonate degradation pathway

nadZ: *gntKU*
- *gntK*: enzyme in idonate degradation pathway
- *gntU*: GNT-transp. (gluconate)



Notes:
- There are multiple homologies between the *idn* and *gnt* genes: *idnT* and *gntU* are homologs and so are *idnK* and *gntK*. Moreover, *idnR* is homolog of the TF encoded in one of its two X-operons, *gntR*, which is located upstream of *gntKU* in the genome. The additional X-operon is CRP

Y-op: *nagBACD*
- *nagBA*: enzymes in N-acetylglucosamine degradation pathway
- *nagC*: TF
- *nagD*: ribonucleotide monophosphatase

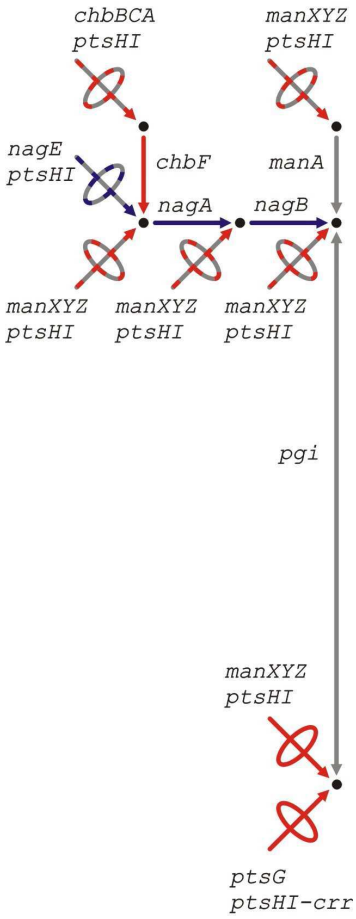
adZ(d): *nagE*
- *nagE* (+ *ptsHI*): PTS-transp. (N-acetylglucosamine)

nadZ: *manXYZ*
- *manXYZ* (+ *ptsHI*) : PTS-transporter (hexoses as N-acetylglucosamine)

nadZ: *chbBCARFG*
- *chbBCA* + (*ptsHI*): PTS-transporter (chitobiose)
- *chbR*: TF
- *chbF*: enzyme in chitobiose degradation pathway
- *chbG*: uncharacterized

nadZ: *nanC-yjhT*
- *nanC*: OmpG-channel(N-acetylneuraminic acid)
- *yjhT*: uncharacterized

Notes:
- *chbBCARFG* is an autoregulated operon
- see Notes for *dgsA* system



Y-op: *dgsA*
- *dgsA*: TF

nadZ: *malT*
- *malT*: TF

nadZ: *manXYZ*
- *manXYZ* (+ *ptsHI*) : PTS-transporter (hexoses as glucose)

nadZ: *ptsG*
- *ptsG* (+ *ptsHI-crr*): PTS-transporter (glucose)

nadZ: *ptsHI-crr*
- *ptsHI-crr*: PTS-transporter (non-specific-sugar components)

Notes:

- *dgsA* and *nagBACD* system (above) are very related: *nagC* and *dgsA* are homologs, and so are *nagE* and *ptsG*; pathways encoded in both systems are closely located in the metabolism and they use the same type of transporters (PTS)

Non-autoregulated Y-operon

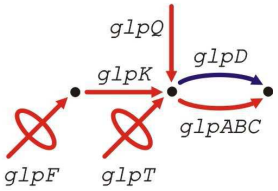
Y-op: *glpEGR*
- *glpE*: thiosulfate sulfurtransferase
- *glpG*: intramembrane serine protease
- *glpR*: TF

adZ(d): *glpD*
- *glpD*: glycerol dehydrogenase (aerobic)

nadZ: *glpABC*
- *glpABC*: glycerol dehydrogenase (anaerobic)

nadZ: *glpTQ*
- *glpT*: MFS-transporter (glycerol-3-P)
- *glpQ*: periplasmic transport associated enzyme

nadZ: *glpFKX*
- *glpF*: MIP-channel (glycerol)
- *glpK*: enzyme for glycerol degradation
- *glpX*: fructose 1,6-bisphosphatase (glycolisis enzyme)

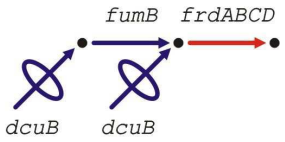


Notes:
- *glpD* and *glpA* are homologs

Y-op: *dcuSR*
- *dcuSR*: 2-component system (anaerobic fumarate respiration)

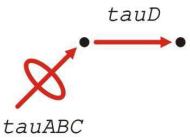
adZ(u): *dcuB-fumB*
- *dcuB*: DCU-transporter (dicarboxylates as fumarate)
- *fumB*: fumarase (anaerobic respiration)

nadZ: *frdABCD*
- *frdABCD*: fumarate reductase (anaerobic respiration)



Y-op: *cbl*
- *cbl*: TF

nadZ: *tauABCD*
- *tauABC*: ABC-transporter (taurine)
- *tauD*: taurine dehydrogenase

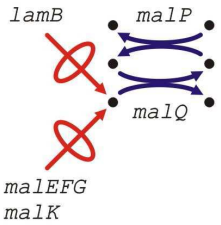


Notes:
- Cbl is homologue of the TF encoded in its only X-operon, *cysB*

Y-op: *malT*
- *malT*: TF

adZ(d): *malPQ*
- *malPQ*: enzymes for maltose and maltodextrins metabolism

nadZ: *malK-lamB-malM*
- *malK* (+ *malEFG*): ABC-transporter (maltose)
- *lamB*: sugar porin (maltose and maltodextrins)
- *malM*: periplasmic protein



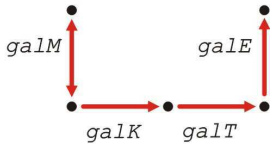
nadZ: *malEFG* (see *malK-lamB-malM*)

nadZ: *malS*
- *malS*: periplasmic maltohexaose transport associated enzyme

Notes:
- *malEFG* and *malK-lamB-malM* are divergent operons: the encoded ABC transporter and porin constitute the maltose/maltodextrin transport system

---

Y-op: *hupA*
- *hupA*: TF



nadZ: *galETKM*
- *galETK*: enzymes for UDP-galactose biosynthesis
- *galM*: galactose-1-epimerase (enzyme that links lactose and galactose metabolisms)

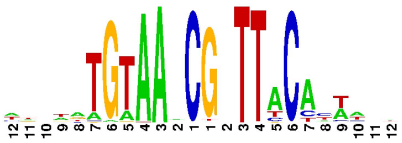
Apéndice D

Logos de los BSs asociados a dominios HTH-LacI

Todos los dominios aquí considerados presentan la secuencia TVSR en el tramo de posiciones que va de AA-17 a AA-20 en la hélice de reconocimiento. Cada logo corresponde al alineamiento de los BSs regulados por TFs con una misma secuencia de aminoácidos en AA-15 y AA-16. Sobre cada logo se muestran la secuencia de los estos aminoácidos de reconocimiento y una tríada de números (i/ii/iii) correspondientes a i) el número total de TFs que tienen esta secuencia, ii) el número de TFs para los que se encontró al menos un BS y iii) el total del BSs encontrados.

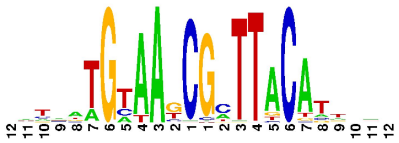
Al pie de cada logo aparecen otras dos tríadas de números que desglosan respectivamente la información concerniente a los procesos de autorregulación (AU) y de regulación del operón codificado corriente abajo en la misma hebra que el operón regulador (arquitectura unidireccional), abreviadamente UD (por, *Unidirectional Downstream*). Cada tríada (i/ii/iii) corresponde ahora a i) el número de secuencias intergénicas de búsqueda; ii) el número de estas secuencias en las que se encontró al menos un BS y iii) el total de BSs encontrados. Nótese que en el caso de la autorregulación el valor de i) coincide con el de TFs pues siempre existe una región intergénica corriente arriba del operón que los codifica. Sin embargo, el operón situado corriente abajo (y la correspondiente zona reguladora que lo antecede) sólo se dispone unidireccionalmente en la mitad de los casos en promedio.

VA 138/93/122



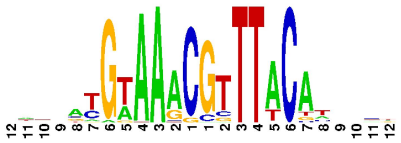
AU: 138/87/103 UD: 67/17/19

TA 101/53/74



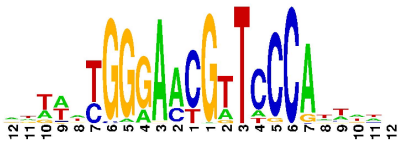
AU: 101/47/63 UD: 61/10/11

IA 91/46/63



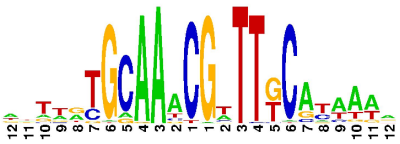
AU: 91/43/55 UD: 45/7/8

KS 87/49/69



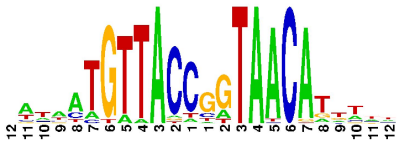
AU: 87/25/33 UD: 55/26/36

PS 76/36/48



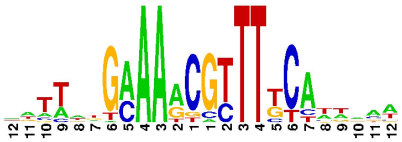
AU: 76/29/39 UD: 44/9/9

KM 68/40/55



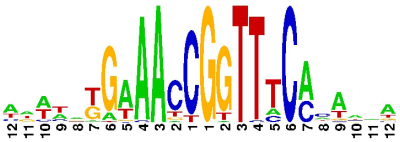
AU: 68/19/20 UD: 39/25/35

IS 57/29/36



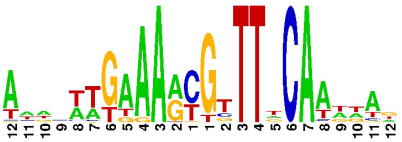
AU: 57/21/25 UD: 28/10/11

KA 50/29/36



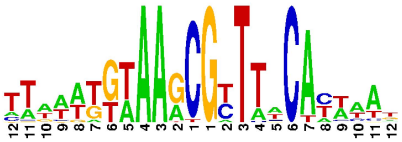
AU: 50/29/30 UD: 24/5/6

MA 47/28/31



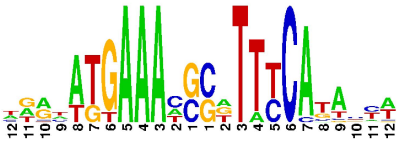
AU: 47/28/31 UD: 26/0/0

LA 44/20/24



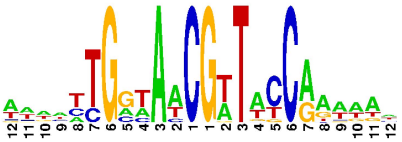
AU: 44/11/12 UD: 38/11/12

VT 40/15/15



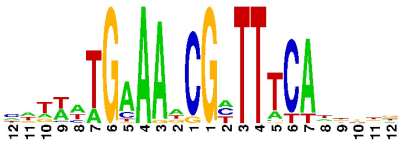
AU: 40/12/12 UD: 26/3/3

VS 36/16/24



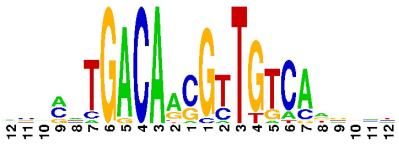
AU: 36/15/22 UD: 21/2/2

PA 34/20/24



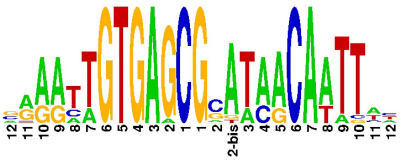
AU: 34/19/19 UD: 23/5/5

IK 29/11/15



AU: 29/10/12 UD: 9/2/3

YQ 24/13/18



AU: 24/0/0 UD: 19/13/18

VG 24/13/20

AU: 24/6/6 UD: 21/7/14

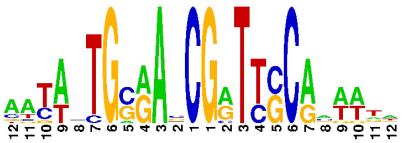
RA 22/0/0

AU: 22/0/0 UD: 8/0/0

QS 22/0/0

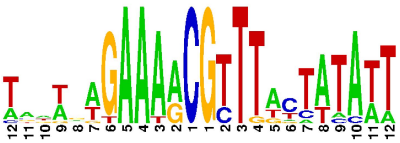
AU: 22/0/0 UD: 9/0/0

RS 21/7/10



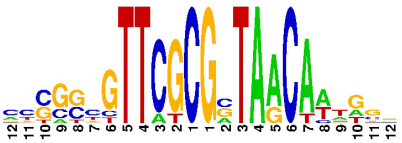
AU: 21/6/6 UD: 18/4/4

PT 21/9/12



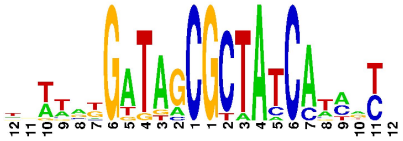
AU: 21/8/10 UD: 6/2/2

EM 20/8/8



AU: 20/8/8 UD: 9/0/0

PM 19/8/11

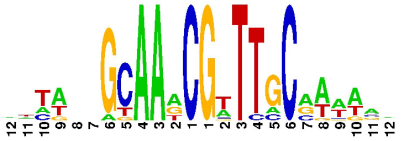


AU: 19/7/9 UD: 5/2/2

QA 18/0/0

AU: 18/0/0 UD: 11/0/0

TS 16/11/13

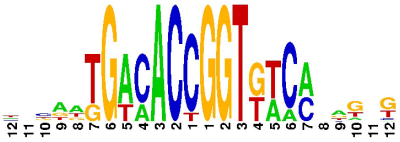


AU: 16/8/8 UD: 11/4/5

LK 14/0/0

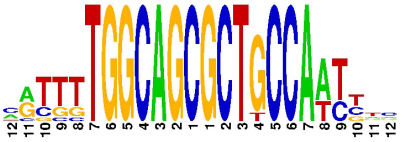
AU: 14/0/0 UD: 10/0/0

KK 14/12/19



AU: 14/5/5 UD: 12/11/14

ES 14/7/9

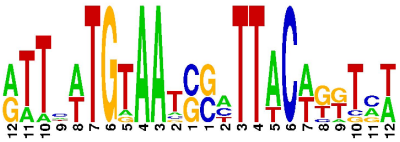


AU: 14/7/9 UD: 5/0/0

AS 14/8/9

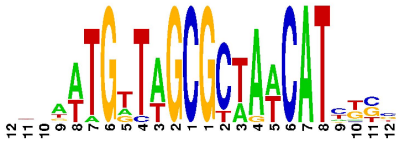
AU: 14/7/8 UD: 8/1/1

SA 13/9/9



AU: 13/9/9 UD: 1/0/0

AI 13/10/13

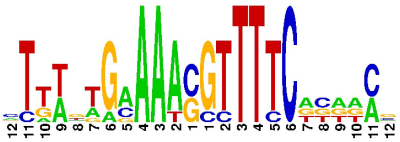


AU: 13/10/12 UD: 4/1/1

TM 12/11/18

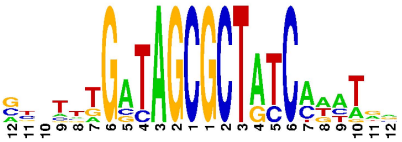
AU: 12/11/18 UD: 5/0/0

IT 12/5/7



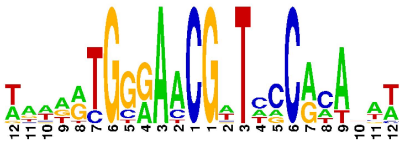
AU: 12/4/4 UD: 9/3/3

PI 11/8/12



AU: 11/8/12 UD: 7/0/0

KT 11/8/11

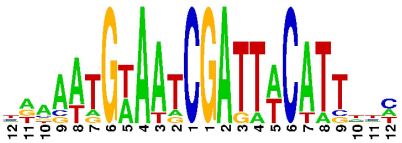


AU: 11/8/8 UD: 6/3/3

AM 10/7/10

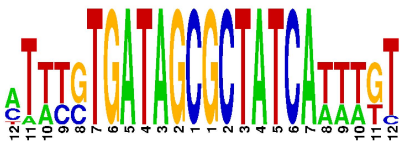
AU: 10/7/9 UD: 5/1/1

AA 10/6/7



AU: 10/5/6 UD: 3/1/1

SM 9/6/9

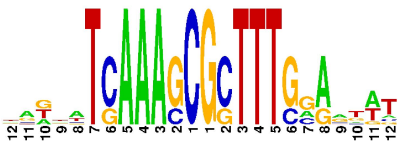


AU: 9/6/9 UD: 0/0/0

RT 9/0/0

AU: 9/0/0 UD: 5/0/0

QT 9/7/8

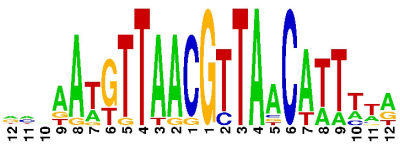


AU: 9/7/8 UD: 7/0/0

QP 9/0/0

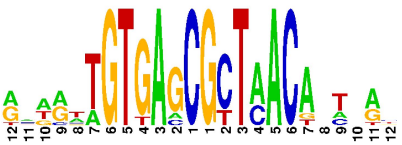
AU: 9/0/0 UD: 4/0/0

LM 9/7/11



AU: 9/5/9 UD: 2/2/2

HQ 9/5/9



AU: 9/2/2 UD: 8/5/7

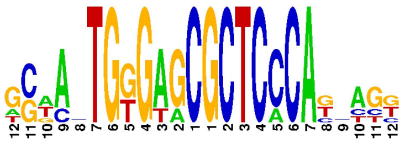
EA 9/0/0

AU: 9/0/0 UD: 8/0/0

RM 8/0/0

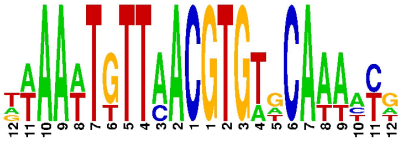
AU: 8/0/0 UD: 5/0/0

RG 8/5/5



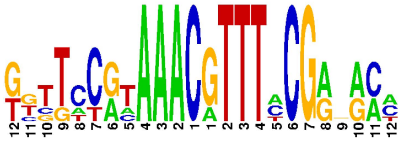
AU: 8/3/3 UD: 5/2/2

HT 8/3/4



AU: 8/2/2 UD: 4/2/2

IG 7/3/5



AU: 7/3/4 UD: 4/1/1

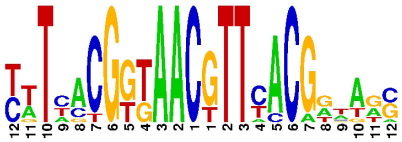
VK 6/0/0

AU: 6/0/0 UD: 4/0/0

LQ 6/5/5

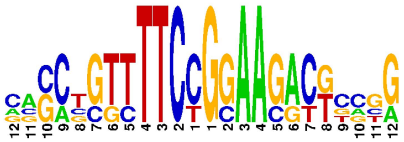
AU: 6/3/3 UD: 3/2/2

IN 6/5/9



AU: 6/4/6 UD: 4/3/3

VP 5/3/4



AU: 5/3/3 UD: 4/1/1

SS 5/0/0

AU: 5/0/0 UD: 2/0/0

TQ 4/0/0

AU: 4/0/0 UD: 0/0/0

TG 4/3/3

AU: 4/3/3 UD: 2/0/0

ST 4/0/0

AU: 4/0/0 UD: 3/0/0

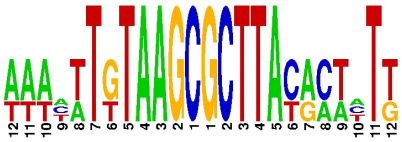
QQ 4/0/0

AU: 4/0/0 UD: 1/0/0

PK 4/0/0

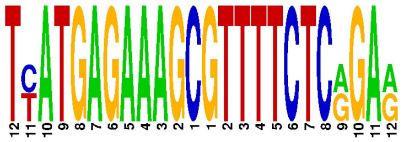
AU: 4/0/0 UD: 4/0/0

IC 4/3/3



AU: 4/2/2 UD: 4/1/1

HA 4/2/2



AU: 4/2/2 UD: 0/0/0

AQ 4/0/0

AU: 4/0/0 UD: 1/0/0

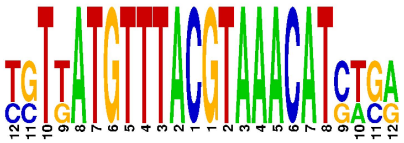
AG 4/0/0

AU: 4/0/0 UD: 2/0/0

YS 3/0/0

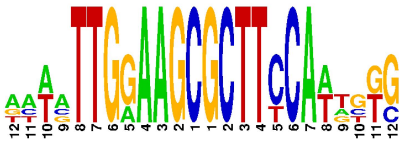
AU: 3/0/0 UD: 3/0/0

SQ 3/2/3



AU: 3/2/3 UD: 3/0/0

RV 3/3/4



AU: 3/1/1 UD: 3/3/3

PQ 3/0/0

AU: 3/0/0 UD: 3/0/0

PG 3/0/0

AU: 3/0/0 UD: 3/0/0

NK 3/3/4

AU: 3/3/4 UD: 2/0/0

NI 3/0/0

AU: 3/0/0 UD: 0/0/0

NA 3/0/0

AU: 3/0/0 UD: 3/0/0

HS 3/2/2

AU: 3/2/2 UD: 1/0/0

HM 3/0/0

AU: 3/0/0 UD: 2/0/0

AT 3/0/0

AU: 3/0/0 UD: 1/0/0